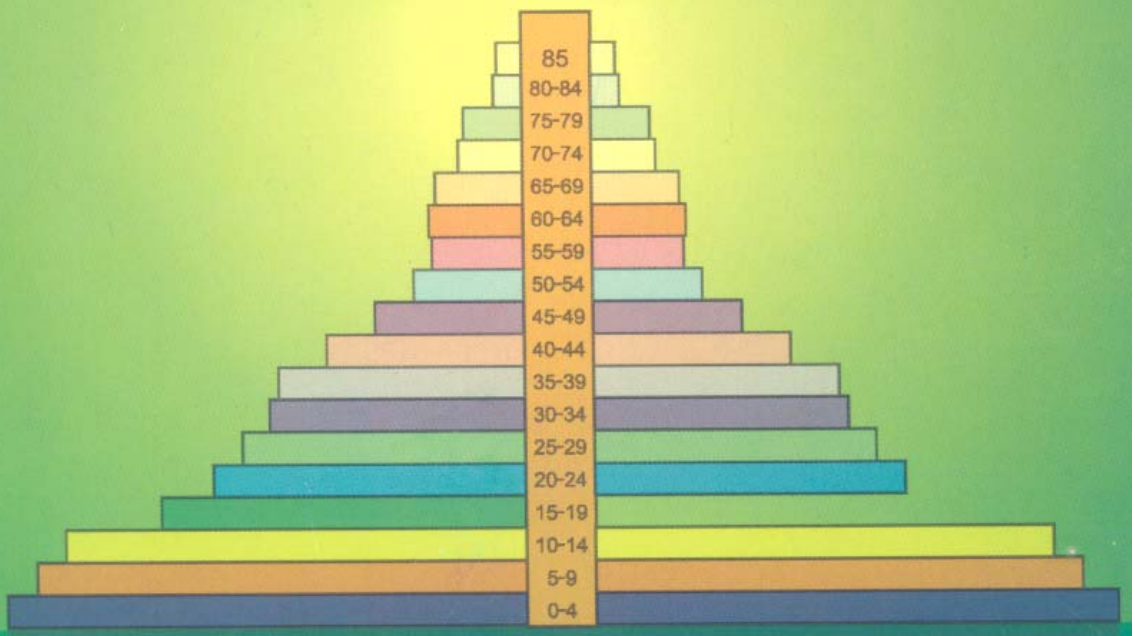
ĐÀO HỮU HỒ

Giáo trình
THỐNG KÊ XÃ HỘI HỌC

DÙNG CHO CÁC TRƯỜNG ĐẠI HỌC KHỐI XÃ HỘI VÀ NHÂN VĂN,
CÁC TRƯỜNG CAO ĐẲNG



NHÀ XUẤT BẢN GIÁO DỤC

ĐÀO HỮU HỒ

GIÁO TRÌNH
THỐNG KÊ XÃ HỘI HỌC

(Dùng cho các trường Đại học khôi Xã hội và Nhân văn, các trường Cao đẳng)

NHÀ XUẤT BẢN GIÁO DỤC

Bản quyền thuộc HEVOBCO – Nhà xuất bản Giáo dục

17–2007/CXB/101–2217/GD Mã số: 7L189M7–DAI

LỜI NÓI ĐẦU

Xác suất - Thống kê là một chuyên ngành khó của Toán học, nhưng lại là chuyên ngành có nhiều ứng dụng trong thực tiễn, và là một trong các công cụ nghiên cứu nhiều chuyên ngành khác. Các chuyên ngành Đại học thuộc khối Xã hội và Nhân văn, cũng như tất cả các trường Cao đẳng, theo Chương trình Khung của Bộ Giáo dục và Đào tạo, đều phải học môn này là một minh chứng rất rõ cho nhận định trên.

Cái khó khi biên soạn giáo trình này không phải là ở nội dung toán học của nó, mà là viết cho đối tượng ít được trang bị về toán, nhất là đối với người học khối Xã hội và Nhân văn. Ngoài kiến thức Toán học ở phổ thông ra, khá nhiều bạn đọc không được trang bị gì thêm về toán cao cấp. Vì vậy, trong giáo trình này Tác giả đã chọn cách trình bày và cố gắng diễn đạt sao cho dễ hiểu nhất đối với bạn đọc. Các khái niệm, các kết quả được trình bày và diễn giải một cách nhẹ nhàng, dễ hiểu, tránh dùng các thuật ngữ, khái niệm trừu tượng, khó hiểu đối với bạn đọc. Việc chứng minh các kết quả cũng được chú ý nhưng ở mức độ vừa phải. Việc giải thích ý nghĩa của khái niệm, ý nghĩa thực tế của bài toán, các bước thực hành cụ thể, v.v... được chú trọng nhiều hơn.

Nội dung chi tiết của giáo trình này phù hợp với nội dung chi tiết môn Thống kê xã hội học hiện đang được giảng dạy trong các trường. Hơn nữa, nội dung chi tiết của giáo trình cũng khá phù hợp với chương trình chi tiết của môn Xác suất thống kê (B) dùng cho các trường Cao đẳng mà Bộ Giáo dục và Đào tạo đã ban hành. Vì vậy, giáo trình này thích hợp và hy vọng là tài liệu có ích cho cả người dạy cũng như người học môn Thống kê xã hội

học ở các trường Đại học khối Xã hội và Nhân văn cũng như môn Xác suất thống kê (B) ở các trường Cao đẳng.

Hiện nay, ở các trường, môn Thống kê xã hội học được giảng dạy với hai mức thời lượng : 45 tiết và 30 tiết. Vì vậy, tác giả cũng biên soạn giáo trình này ở cả hai mức tương ứng. Nếu với thời lượng 45 tiết, bạn đọc hãy dùng Chương I (22 tiết) và Chương II (23 tiết). Nhưng ở mức độ 30 tiết bạn đọc hãy bỏ qua Chương I và thay vào đó là phần Phụ lục I (8 tiết), sau đó là Chương II (22 tiết).

Riêng đối với Chương I, phần biến ngẫu nhiên và các khái niệm liên quan (I.6; I.7; I.8) yêu cầu thực hành chỉ đặt ra đối với biến rời rạc, còn đối với biến liên tục chỉ yêu cầu bạn đọc biết được các khái niệm và công thức tương ứng.

Mặc dù đã cố gắng, song khó tránh khỏi sai sót. Tác giả mong nhận được sự lượng thứ và đóng góp ý kiến của bạn đọc. Mọi ý kiến xin gửi về Ban Biên tập Sách Đại học – Cao đẳng, CTCP Sách Đại học – Dạy nghề, Nhà Xuất bản Giáo dục – 25 Hàn Thuyên – Hà Nội.

Hà Nội, ngày 31/12/2006

TÁC GIẢ

Chương I

MỘT SỐ KHÁI NIỆM VÀ KẾT QUẢ CƠ BẢN CỦA XÁC SUẤT

I.1. GIẢI TÍCH TỔ HỢP

Giải tích tổ hợp, bạn đọc đã được học ở THPT, ngay cả đối với ban Xã hội và Nhân văn. Do đó, phần này chỉ nhắc lại những điểm cần hiểu rõ của khái niệm để tránh nhầm lẫn khi dùng. Ví dụ minh họa được kết hợp trong các bài toán tính xác suất ở mục sau. Thực ra, trong các kết quả của giải tích tổ hợp, với mức độ của giáo trình này, chỉ yêu cầu bạn đọc hiểu và dùng được tổ hợp, luật tích. Hoán vị, chỉnh hợp, chỉnh hợp lặp có thể được suy ra từ tổ hợp và luật tích (xem [1]).

Tổ hợp:

Khi lấy ngẫu nhiên ra k phần tử từ một tập gồm n phần tử (ở đây là lấy đồng thời, lấy cùng lúc, lấy một lần ra k phần tử; $k \leq n$), sao cho hai cách lấy ra k phần tử được gọi là khác nhau nếu giữa chúng có ít nhất một phần tử khác nhau (nghĩa là sự khác nhau về thứ tự của các phần tử không có ý nghĩa gì đối với cách lấy theo tổ hợp) thì: số cách lấy ra k phần tử

từ n phần tử như trên được gọi là tổ hợp chập k của n, được ký hiệu là C_n^k.

Tổ hợp này được xác định như sau:

$$C_n^k = \frac{n!}{k!(n-k)!}$$

trong đó: $n! = n.(n-1).(n-2)... 3.2.1 = n.(n-1)! = n.(n-1).(n-2)!$

$= n.(n-1) ... (n-(k-1)).(n-k)!$

$0! = 1$

Chữ C là viết tắt của từ combination, nghĩa là tổ hợp.

Rõ ràng, ta thấy C_n^k phải là số nguyên, dương.

Ta có: $C_n^k = C_n^{n-k}$

$C_n^0 = C_n^n = 1$

$C_n^1 = C_n^{n-1} = n$

(Bạn đọc hãy làm quen với việc tính nhẩm: C_6^1, C_3^2, C_4^2, C_6^2, C_6^3, C_{10}^3, C_{10}^7, ...)

Trong máy tính Casio bỏ túi cũng đã có chương trình để tính C_n^k).

Luật tích:

Giả sử hiện tượng A được thực hiện bởi k bước liên tiếp (k = 2, 3, ...), trong đó bước thứ i có n_i cách thực hiện. Khi đó, để nhận được A ta có $(n_1.n_2 ... n_k)$ cách thực hiện.

Luật tổng:

Giả sử hiện tượng A được thực hiện nếu B được thực hiện hoặc nếu C được thực hiện. Khi đó, để nhận được A ta có $(n_B + n_C)$ cách thực hiện với n_B, n_C là số cách thực hiện B và C tương ứng.

I.2. PHÉP THỬ VÀ BIẾN CỐ

Trước hết, chúng ta bắt đầu từ những phép thử quen thuộc:

Gieo một đồng tiền trên một mặt phẳng. Đó là một phép thử. Phép thử này có hai khả năng (tình huống) có thể xảy ra, đó là "xuất hiện mặt sấp" và "xuất hiện mặt ngửa". Đấy cũng là hai biến cố sơ cấp.

Gieo một con xúc xắc trên mặt phẳng. Đó là một phép thử. Phép thử này có 6 khả năng (tình huống) có thể xảy ra. Đó là "xuất hiện k chấm ở mặt trên của con xúc xắc", k = 1, 2..., 6. Đó cũng là 6 biến cố sơ cấp. Nhưng tình huống "xuất hiện mặt có số chấm chẵn" sẽ chỉ là biến cố, không phải là biến cố sơ cấp. Rõ ràng, "xuất hiện mặt có số chấm chẵn" cũng là một tình huống của phép thử. Vậy biến cố và biến cố sơ cấp khác nhau ở điểm nào?

Chọn ngẫu nhiên một đại biểu, phỏng vấn ngẫu nhiên một khách hàng,... Đó cũng là các phép thử. Tùy yêu cầu của phép thử mà ta có các khả năng có thể khác nhau. Chẳng hạn, xét về giới tính của đại biểu thì phép thử có hai khả năng có thể, nhưng xét về thành phần giai cấp, xét về dân tộc, xét về nghề nghiệp,... thì phép thử lại có nhiều khả năng có thể.

Bắn một viên đạn vào một mục tiêu cũng là một phép thử. Phép thử này có hai khả năng: có thể "trúng mục tiêu" và "không trúng mục tiêu", cùng là hai biến số sơ cấp. Bắn một viên đạn vào bia để tính điểm – phép thử có 11 khả năng có thể: "Bắn được k điểm", k = 0, 1, ..., 10. Đó cũng là 11 biến cố sơ cấp. Nhưng "Bắn được điểm giỏi" không phải là biến cố sơ cấp.

Qua các ví dụ trên, chúng ta cần hình thành một số khái niệm: phép thử, biến cố, biến cố sơ cấp.

– Thực hiện một hành động nào đó tức là ta đã thực hiện một phép thử. Phép thử mà ta không khẳng định chắc chắn được kết quả trước khi nó được thực hiện gọi là phép thử ngẫu nhiên.

– Một khả năng (tình huống) có thể xảy ra của phép thử được gọi là biến cố.

– Biến cố không phân tích nhỏ hơn được nữa được gọi là biến cố sơ cấp.

Lưu ý rằng số biến cố sơ cấp sẽ phụ thuộc vào nội dung và yêu cầu của phép thử, chứ không phụ thuộc vào người thực hiện phép thử.

Các biến cố được phân chia thành ba loại chính như sau:

– Biến cố không thể, ký hiệu ϕ, là biến cố không thể xảy ra khi phép thử được thực hiện.

– Biến cố chắc chắn, ký hiệu Ω, là biến cố nhất định xảy ra khi phép thử được thực hiện.

– Biến cố ngẫu nhiên, ký hiệu A, B, C,..., là biến cố có thể xảy ra và cũng có thể không xảy ra khi phép thử được thực hiện.

Nghiên cứu các phép thử ngẫu nhiên, tức là nghiên cứu các kết quả có thể của phép thử, nghĩa là nghiên cứu các biến cố ngẫu nhiên chính là đối tượng nghiên cứu đầu tiên của Lý thuyết Xác suất.

I.3. ĐỊNH NGHĨA XÁC SUẤT

Trong triết học cũng có phạm trù tất nhiên và ngẫu nhiên – Hiện tượng ngẫu nhiên trong triết học cũng được hiểu tương tự như biến cố ngẫu nhiên nói trên. Nhưng cách

nghiên cứu tính ngẫu nhiên trong triết học khác xa với cách nghiên cứu tính ngẫu nhiên của toán. Để nghiên cứu các biến cố ngẫu nhiên, các nhà toán học đã xây dựng một khái niệm mới, được gọi là xác suất. Ở mức độ đơn giản dưới đây chỉ nêu định nghĩa xác suất dạng cổ điển.

Định nghĩa:

Xác suất của một biến cố A là một số không âm, ký hiệu là P(A), biểu thị khả năng xảy ra biến cố A. P(A) được xác định như sau:

$$P(A) = \frac{\text{Số biến cố sơ cấp thuận lợi cho A}}{\text{Số biến cố sơ cấp của phép thử}}$$

Chữ P là viết tắt của từ probability, nghĩa là xác suất.

Biến cố sơ cấp được gọi là thuận lợi cho biến cố A nếu nó xảy ra thì suy ra biến cố A xảy ra. Định nghĩa này đúng với điều kiện các biến cố sơ cấp có cùng khả năng xảy ra, do đó người ta gọi định nghĩa này là định nghĩa xác suất theo tính đồng khả năng.

Tính chất của xác suất:

$0 \le P(A) \le 1 = 100\%$

$P(\phi) = 0$; $P(\Omega) = 1$

$P(A) + P(\bar{A}) = 1$ ($\bar{A}$ được tạm hiểu là phủ định của A).

Xác suất là một khái niệm mới, nhưng thực chất lại là một khái niệm rất quen thuộc. Đó là khả năng xảy ra. Suy nghĩ về khả năng xảy ra chúng ta sẽ thấy các yêu cầu, các tính chất của xác suất được nêu ở trên là hợp lý và đúng đắn. Như vậy, bạn đọc đã tự cho mình một cách chứng minh đơn giản.

Ở phần trên có đề cập đến số khả năng. Số khả năng khác với khả năng xảy ra mà ta dùng để diễn đạt ý nghĩa của xác suất.

Nhận xét: Theo định nghĩa cổ điển, để tìm xác suất P(A) ta sẽ tìm hai con số ở tử số và mẫu số, rồi làm phép chia. Việc tìm hai con số trên lại là bài toán sơ cấp: dùng giải tích tổ hợp hoặc đếm trực tiếp. Thông thường, chúng ta tìm số biến cố sơ cấp của phép thử trước, mà muốn tìm con số này dễ dàng thì ta phải phân tích phép thử để xem phép thử thực hiện một lần (lấy theo nghĩa tổ hợp) hay thực hiện k lần (dùng luật tích), sau đó tìm số biến cố sơ cấp thuận lợi cho A. (Bạn đọc cần phân biệt số lần thực hiện phép thử với số cách thực hiện phép thử).

Ví dụ I.1: Từ một tổ gồm 9 nam và 6 nữ, ta chọn ngẫu nhiên ra 5 người. Tìm xác suất:

a) Chọn được 3 nam.

b) Chọn được 3 nữ.

c) Chọn được 1 nam.

d) Chọn được 5 nữ.

e) Chọn được 5 nam.

Giải:

Phép thử ở đây là lấy cùng lúc ra 5 người (lấy một lần). Do đó, số cách lấy sẽ là $C_{15}^5 = 3003$ cách, hay ta có 3003 biến cố sơ cấp ứng với phép thử đang xét.

Đặt A = {Chọn được 3 nam trong 5 người chọn ra} = {Chọn được 3 nam và 2 nữ}. Tương tự đối với các biến cố B, C, D, E.

a) Để được A phải chọn hai lần, đầu tiên chọn ra 3 nam trong số 9 nam, sau đó chọn ra 2 nữ trong số 6 nữ. Theo luật tích ta có số cách chọn thuận lợi cho A là:

$$C_9^3 . C_6^2 = 84.15 = 1260$$

Vậy: $P(A) = \frac{1260}{3003} = 0,4196$

b) Tương tự, số biến cố sơ cấp thuận lợi cho B là:

$C_9^2 . C_6^3 = 720 \rightarrow P(B) = \frac{720}{3003} = 0,2398$

c) $P(C) = \frac{C_9^1 . C_6^4}{C_{15}^5} = \frac{135}{3003} = 0,0450$

d) $P(D) = \frac{C_9^0 . C_6^5}{C_{15}^5} = \frac{6}{3003} = 0,0020$

e) $P(E) = \frac{C_9^5 . C_6^0}{C_{15}^5} = \frac{126}{3003} = 0,0420$

Qua 5 xác suất tìm được, ta thấy khả năng xảy ra biến cố A là cao nhất (~ 42%), còn khả năng xảy ra biến cố D là nhỏ nhất (~ 0,2%). Khi thực hiện chọn ra 5 người cùng lúc trong một lần nào đó thì trong 5 biến cố trên, chúng ta trông chờ xảy ra biến cố A hoặc B và không hy vọng xảy ra biến cố C, D, E. Nhưng nếu thực hiện phép thử trên 1000 lần, tức là 1000 lần lấy ra 5 người từ 15 người này thì sẽ có khoảng 420 (~ 419,6) lần xảy ra biến cố A, khoảng 240 (~ 239,8) lần xảy ra biến cố B, khoảng 45 lần xảy ra biến cố C, khoảng 42 lần xảy ra biến cố E và chỉ khoảng 2 lần xảy ra biến cố D. (Bạn đọc sẽ hiểu điều này hơn khi học đến dãy phép thử Bernoulli).

Ví dụ I.2: Xét hai tình huống sau:

a) Một đại hội gồm 100 đại biểu, trong đó có 30 đại biểu nữ. Chọn ngẫu nhiên ra một đại biểu. Tìm xác suất chọn được đại biểu nữ.

b) Tỷ lệ học sinh giỏi của trường là 20%. Trong lúc học

sinh đang tập trung ở sân trường để sinh hoạt chung, hãy chọn ngẫu nhiên một học sinh. Cho biết xác suất chọn được học sinh giỏi.

Giải:

a) P (Chọn được đại biểu nữ) = $\frac{C_{30}^1}{C_{100}^1} = \frac{30}{100}$

Phân số $\frac{30}{100}$ lại chính là tỷ lệ đại biểu nữ của đại hội.

b) P (Chọn được học sinh giỏi) = 20% = tỷ lệ học sinh giỏi của trường.

Như vậy, *xác suất lại chính là một tỷ lệ nào đó*. Tỷ lệ là một đại lượng rất quen thuộc, được dùng rộng rãi và chúng ta đã được học tính tỷ lệ ngay từ bậc Tiểu học.

Bạn đọc hãy diễn đạt tương đương theo chiều ngược lại các mệnh đề sau: Tỷ lệ phiếu bầu cho ứng cử viên A là 42%. Tỷ lệ các cô gái cao trên 1,60m là 22%. Tỷ lệ các hộ gia đình có thu nhập từ 5 triệu đồng đến 8 triệu đồng ở thành phố Hà Nội là 45%. Đó chính là: Xác suất chọn được một cử tri bầu cho ứng cử viên A là 42%. Xác suất chọn được cô gái cao trên 1.60m là 22%. Xác suất chọn được hộ gia đình ở Hà Nội có thu nhập từ 5 triệu đến 8 triệu đồng là 45%.

Ví dụ I.3: Giả sử có 10 mảnh bìa vuông như nhau, được ghi các số từ 0 đến 9. Ta rút ngẫu nhiên một bìa và ghi lại số trên bìa đó (ký hiệu là X). Trả bìa đó trở lại tập ban đầu, xáo trộn đều, sau đó lại rút hú họa ra một bìa, ghi lại số của nó (ký hiệu là Y). Hỏi:

a) Có bao nhiêu biến cố sơ cấp của phép thử trên?

b) Tính P (X = 3)

c) Tính P ($X \leq 3$).

d) Tính P ($X > 3$ và $Y > 3$).

e) Tính P ($X + Y < 6$).

f) Tính P ($X \neq 3$).

g) Tính P (X hoặc $Y = 5$).

h) Tính P (X hoặc $Y \leq 6$).

i) Tính P ($X = Y$).

j) Tính P ($X = 5$ và $Y \neq 5$).

k) Tính P ($4 < X < 6$ và $4 < Y < 6$).

l) Tính P ($X - Y = 2$).

Giải:

a) Phép thử thực hiện 2 lần. Theo luật tích, ta có số biến cố sơ cấp là $C_{10}^1 . C_{10}^1 = 100$.

b) $P(X = 3) = \frac{1}{10} = 10\%$

Có 2 cách tính như sau:

$P(X = 3) = P(X = 3, Y \text{ tùy ý}) = \frac{1.10}{100} = \frac{1}{10}$

Hoặc chỉ xét phép thử liên quan đến X, khi đó số biến cố sơ cấp là 10, số thuận lợi là 1.

c) $P(X \leq 3) = \frac{4}{10} = 40\%$

d) $P(X > 3 \text{ và } Y > 3) = \frac{6.6}{100} = 36\%$

e) $P(X + Y < 6) = \frac{6 + 5 + 4 + ... + 1}{100} = 21\%$

(Nếu $X = 0$ thì Y có thể từ 0 đến 5: có 6 trường hợp,

Nếu X = 1 thì Y có thể từ 0 đến 4: có 5 trường hợp, v.v...)

f) $P(X \neq 3) = \frac{9}{10} = 90\%$

g) $P(X \text{ hoặc } Y = 5) = \frac{10+9}{100} = \frac{19}{100} = 19\%$

(Nếu X = 5, Y có 10 khả năng. Nếu Y = 5 thì X còn 9 khả năng (trừ khả năng Y = 5, X = 5 đã tính trước rồi).

h) $P(X \text{ hoặc } Y \leq 6) = \frac{70+21}{100} = 91\%$

(Nếu X = 0, Y có 10 khả năng, ...

Nếu X = 6, Y có 10 khả năng.

Nếu Y = 0, X chỉ còn 3 khả năng là 7 hoặc 8 hoặc 9,...

Nếu Y = 6 thì X chỉ còn 3 khả năng.

Vậy số thuận lợi là: 10.7 + 3.7 = 91)

i) $P(X = Y) = \frac{10}{100} = 10\%$

j) $P(X = 5 \text{ và } Y \neq 5) = \frac{9}{100} = 9\%$

k) $P(4 < X < 6 \text{ và } 4 < Y < 6) = \frac{1}{100} = 1\%$

l) $P(X - Y = 2) = \frac{8}{100} = 8\%$

Ví dụ I.4: Đoàn thanh niên tổ chức vui chơi, kết hợp quay số có thưởng. Ban tổ chức đã phát ra 1000 vé (được đánh số từ 000 đến 999). Tìm xác suất để khi quay giải nhất ta nhận được:

a) Vé có chữ số hàng đơn vị chẵn.

b) Vé có 3 chữ số khác nhau.

c) Vé có chữ số đầu là 8 và 2 chữ số còn lại khác nhau.

d) Vé có 3 chữ số trùng nhau.

Giải:

Phép thử ở đây là 3 lần quay (3 lần chọn), mỗi lần chọn một chữ số trong 10 chữ số từ 0 đến 9. Do đó, số biến cố sơ cấp = 10.10.10 = 1000.

Đặt A = {Quay được vé có chữ số hàng đơn vị chẵn}. Tương tự cho B, C, D.

a) $P(A) = \dfrac{C_{10}^1 . C_{10}^1 . C_5^1}{10^3} = 0,50$

b) $P(B) = \dfrac{C_{10}^1 . C_{10}^1 . C_5^1}{10^3} = \dfrac{10.9.8}{1000} = 0,720$

(Số biến cố sơ cấp thuận lợi cho B: 3 lần chọn, lần I có 10 cách chọn, lần II phải bớt đi chữ số đã chọn nên chỉ còn $C_9^1 = 9$ cách chọn, lần III phải bớt đi 2 chữ số đã chọn nên chỉ còn $C_8^1 = 8$ cách chọn. Theo luật tích, số cách chọn thuận lợi cho B là 10.9.8 – Mà tích này cũng chính là chỉnh hợp. Như vậy, dùng tổ hợp và luật tích ta cũng nhận được kết quả của chỉnh hợp).

c) $P(C) = \dfrac{1.C_{10}^1 . C_9^1}{10^3} = \dfrac{1.10.9}{1000} = 0,09$

d) $P(D) = \dfrac{C_{10}^1 . 1.1}{10^3} = 0,01$

Qua các ví dụ trên giúp chúng ta hiểu rõ hơn nhận xét đã nêu trước ví dụ I.1 (trang 12). Đó chính là chìa khóa để giải các bài toán tính xác suất bằng định nghĩa cổ điển. Nắm vững chìa khóa này bạn đọc có thể giải được các bài toán xác suất bằng định nghĩa cổ điển ở mức khó, vượt xa yêu cầu của giáo trình này.

I.4. CÔNG THỨC XÁC SUẤT CỦA TỔNG VÀ TÍCH HAI BIẾN CỐ

Để mở rộng việc tính xác suất của một biến cố người ta đã xây dựng các phép toán đối với các biến cố. Có thể nói các phép toán này được hiểu gần tương tự như các phép toán tập hợp mà bạn đọc đã được biết ở THPT. Dưới đây nhắc lại một vài khái niệm cần dùng.

– Biến cố A kéo theo biến cố B, ký hiệu $A \subset B$, nếu A xảy ra thì suy ra B xảy ra.

– Hai biến cố A và B tương đương, ký hiệu $A = B$, nếu $A \subset B$ và $B \subset A$.

– Tổng của 2 biến cố A và B là biến cố tổng $A \cup B$ sao cho $A \cup B$ xảy ra khi và chỉ khi hoặc A xảy ra, hoặc B xảy ra.

Ta có: $A \cup B$ xảy ra tương đương với biến cố {có ít nhất một biến cố xảy ra}.

– Tích của 2 biến cố A và B là biến cố tích AB sao cho AB xảy ra khi và chỉ khi A xảy ra và B xảy ra.

– Hai biến cố xung khắc nếu tích của chúng tương đương với biến cố không thể hoặc nếu biến cố này xảy ra thì biến cố kia không xảy ra.

– Biến cố đối lập; $\bar{A}$ được gọi là biến cố đối lập của A nếu A, $\bar{A}$ xung khắc ($A\,\bar{A} = \Phi$) ; $\bar{A} \cup A = \Omega$.

– Hai biến cố A và B độc lập với nhau nếu việc xảy ra biến cố này hay không, không ảnh hưởng đến khả năng xảy ra biến cố kia.

Dễ thấy rằng: Nếu A, B độc lập với nhau thì A, $\bar{B}$ hoặc $\bar{A}$, B hoặc $\bar{A}$, $\bar{B}$ cũng độc lập với nhau (Do A độc lập với B nên việc A xảy ra hay không, không ảnh hưởng đến P(B), vì

vậy cũng không ảnh hưởng đến $1 - P(B) = P(\overline{B})$, nghĩa là A độc lập với $\overline{B}$, v.v...).

Biểu diễn hình học các khái niệm trên là khó bởi vì biến cố là mệnh đề logic. Song các khái niệm trên có sự tương tự như các khái niệm của tập hợp, vì thế người ta mượn sơ đồ Venn (tên của nhà logic người Anh – John Venn) biểu diễn các khái niệm của tập hợp để biểu diễn các khái niệm của biến cố. Đối với sơ đồ Venn bạn đọc đã quen biết nên trong phần này không trình bày lại (xem [1], [2]). Nhưng dùng sơ đồ Venn để hiểu biến cố tích như là phần chung của hai biến cố, như là phần mà A, B cùng xảy ra,... thì lại là không đúng.

Bây giờ chúng ta xây dựng công thức xác suất của tổng và tích hai biến cố.

Ta có:

$$P(A \cup B) = P(A) + P(B) - P(AB) \qquad (I.1)$$

Nếu A, B xung khắc:

$$P(A \cup B) = P(A) + P(B) \qquad (I.2)$$

Theo (I.2), ta có tính chất quen thuộc của xác suất:

$$\left.\begin{aligned} P(A \cup \overline{A}) &= P(\Omega) = 1 \\ P(A \cup \overline{A}) &= P(A) + P(\overline{A}) \end{aligned}\right\} \Rightarrow P(A) + P(\overline{A}) = 1$$

Nếu A, B độc lập:

$$P(AB) = P(A).\ P(B) \qquad (I.3)$$

Phép tổng và tích hai biến cố hoàn toàn được mở rộng cho ba, bốn,... biến cố. Công thức (I.1), (I.2), (I.3) cũng được xây dựng cho ba, bốn,... biến cố. Song ở mức độ đơn giản của giáo trình, chúng ta dừng lại ở đây.

Để chứng minh hai công thức (I.1) và (I.3), chúng ta chứng minh đại diện, chẳng hạn chứng minh công thức (I.1).

Gọi n là số biến cố sơ cấp của phép thử.

Gọi m_A, m_B, m_{AB} là số biến cố sơ cấp thuận lợi cho A, B, AB tương ứng. Khi đó, số biến cố sơ cấp thuận lợi cho biến cố

tổng $A \cup B$ sẽ là: $m_A + m_B - m_{AB}$ (bằng số biến cố sơ cấp thuận lợi cho A cộng với số biến cố sơ cấp thuận lợi cho B nhưng phải trừ đi số biến cố sơ cấp thuận lợi cho AB vì số này đã được kể đến trong số m_A).

Theo định nghĩa cổ điển ta có:

$$P(A \cup B) = \frac{m_A + m_B - m_{AB}}{n} = \frac{m_A}{n} + \frac{m_B}{n} - \frac{m_{AB}}{n}$$

$$= P(A) + P(B) - P(AB).$$

Công thức (I.1) được chứng minh.

Ví dụ I.5: Hai vận động viên A và B của địa phương Z tham gia giải bóng bàn đơn nữ toàn quốc. Khả năng lọt qua vòng loại để vào vòng chung kết của từng người tương ứng là 80% và 60% (mỗi bảng chỉ chọn một người vào vòng chung kết và hai vận động viên A, B không cùng trong một bảng đấu loại). Tìm khả năng xảy ra các tình huống sau:

a) Cả hai lọt vào vòng chung kết.

b) Có ít nhất một người lọt vào vòng chung kết.

c) Chỉ có vận động viên A lọt vào vòng chung kết.

Giải:

Rõ ràng với điều kiện đã cho, bài toán này không thể tính xác suất bằng định nghĩa cổ điển được.

Bài toán cho hai biến cố:

Đặt A = {Vận động viên A lọt vào vòng chung kết}

B = {Vận động viên B lọt vào vòng chung kết}

A, B độc lập, không xung khắc. Theo đề bài ta có P(A) = 0,80; P(B) = 0,60.

a) Đặt A_1 = {Cả hai vận động viên lọt vào vòng chung kết}
= A.B

$\Rightarrow P(A_1) = P(AB) = P(A).P(B) = 0,8.0,6 = 0,48$

b) Đặt A_2 = {Có ít nhất một người lọt vào vòng chung kết}
$= A \cup B$

$\Rightarrow P(A_2) = P(A \cup B) = P(A) + P(B) - P(AB)$
$= 0,80 + 0,60 - 0,48 = 0,92$

c) Đặt A_3 = {Chỉ có A lọt vào vòng chung kết} = $A.\overline{B}$

$\Rightarrow P(A_3) = P(A\overline{B}) = P(A).P(\overline{B})$
$= 0,80.(1 - 0,6) = 0,32.$

Qua ba xác suất tính được, ta thấy tình huống b) là có khả năng xảy ra cao nhất. Tức là địa phương Z có cơ sở để trông chờ kết quả này.

Nhận xét: Loại đơn giản của mô hình này là bài toán đã cho một vài xác suất, nghĩa là đã có một vài biến cố đã cho. Vì vậy, trước tiên phải đặt tên các biến cố đã cho, nhận xét tính xung khắc, độc lập của chúng. Sau đó, biểu diễn biến cố cần tìm xác suất qua các biến cố đã cho (kể cả các biến cố đối lập). Đây là bước khó nhất của mô hình này. Vì chỉ có hai công thức, xác suất của biến cố tổng và xác suất của biến cố tích nên chúng ta chỉ quan tâm đến hai cách biểu diễn: tổng và tích của những biến cố đã cho. Một dấu hiệu đơn giản là: Khi diễn đạt thành lời biến cố cần tìm xác suất, nếu chúng ta dùng từ "hoặc" thì nên nghĩ ngay đến phép tổng, còn nếu dùng từ "và" thì nên nghĩ về phép tích.

Bước thứ ba phải làm là áp dụng công thức (I.1) hoặc (I.2) hoặc (I.3) để tính các xác suất cần tìm.

Bạn đọc hãy vận dụng nhận xét này với ví dụ I.5 ở trên.

{Cả hai vận động viên lọt vào chung kết} = {A lọt vào chung kết và B lọt vào chung kết} = A.B

{Có ít nhất một người lọt vào chung kết } = {Hoặc A lọt vào chung kết hoặc B lọt vào chung kết } = $A \cup B$

(Biến cố này tương đương với biến cố tổng, như đã nêu ở trên)

{Chỉ có A lọt vào chung kết } = {A lọt vào chung kết và B không lot vào chung kết } = $A.\overline{B}$.

I.5. DÃY PHÉP THỬ BERNOULLI

I.5.1. Định nghĩa

– Hai phép thử được gọi là độc lập với nhau nếu việc thực hiện và kết quả của phép thử này độc lập và không ảnh hưởng đến việc thực hiện và kết quả của phép thử kia.

– n phép thử độc lập được gọi là n phép thử Bernoulli nếu thỏa mãn:

a) Mỗi phép thử xảy ra một trong hai biến cố là A và $\bar{A}$.

b) Khả năng xảy ra biến cố A là như nhau đối với mọi phép thử:

$P(A) = p \Rightarrow P(\bar{A}) = 1 - p$

n phép thử độc lập thỏa mãn hai điều kiện trên gặp rất nhiều trong thực tế, để đơn giản ta gọi là dãy Bernoulli (tên của nhà bác học đưa ra định nghĩa này). Chẳng hạn, gieo một đồng tiền n lần sẽ là n phép thử Bernoulli. Biến cố A có thể là: {Xuất hiện mặt sấp}. Xác suất p = P(A) sẽ là 1/2 nếu đồng tiền cân đối, đồng chất. Gieo một con xúc xắc 10 lần sẽ là 10 phép thử Bernoulli; biến cố A có thể là {Xuất hiện mặt lục} hoặc {Xuất hiện mặt có số chấm chẵn},... tùy theo yêu cầu của bài toán muốn quan tâm đến tình huống nào. Nếu con xúc xắc cân đối, đồng chất thì ta tìm ngay được xác suất p = P(A). Một xạ thủ bắn 60 viên đạn vào bia để tính điểm (tất nhiên với cùng một khẩu súng và cùng một loại đạn). Đó cũng là 60 phép thử Bernoulli. Biến cố A có thể là {Bắn được 10 điểm} hoặc {Bắn đạt điểm giỏi} hoặc {Bắn đạt yêu cầu},... Tùy theo yêu cầu của bài toán mà ta xác định biến cố A v.v...

I.5.2. Xác suất P_n (m; p)

Khi thực hiện n phép thử Bernoulli thì biến cố A có thể xảy ra 0 lần, 1 lần, ... và n lần. Biến cố {Trong n phép thử Bernoulli biến cố A xuất hiện m lần} là một biến cố ngẫu nhiên. Xác suất của biến cố này được ký hiệu là $P_n(m; p)$.

Ta có:

$$P_n(m; p) = C_n^m p^m (1-p)^{n-m} \; ; m = 0, 1, 2, ..., n. \qquad (I.4)$$

Để chứng minh công thức (I.4) ta xét ví dụ cụ thể sau: Gieo một con xúc xắc cân đối, đồng chất 10 lần. Tìm xác suất để có 3 lần xuất hiện mặt lục.

Ta có 10 phép thử Bernoulli với A = {Xuất hiện mặt lục} và p = P(A) = 1/6. Giả sử lần gieo thứ nhất, thứ tư và thứ chín xuất hiện mặt lục. Khi đó, kết quả của 10 lần gieo này có thể biểu diễn là: $A\ \bar{A}\ \bar{A}\ A\ \bar{A}\ \bar{A}\ \bar{A}\ \bar{A}\ A\bar{A}$

Do các lần gieo độc lập, các biến cố trong dãy trên là độc lập, nên:

$$P(A\bar{A}\ \bar{A}\ A\bar{A}\ \bar{A}\ \bar{A}\ \bar{A}\ A\bar{A}) =$$

$$P(A).P(\bar{A}).P(\bar{A}).P(A).P(\bar{A}).P(\bar{A}).P(\bar{A}).P(\bar{A}).P(A).P(\bar{A})$$

$$= p(1-p)(1-p)p(1-p)(1-p)(1-p)(1-p)p(1-p)$$

$$= p^3(1-p)^7 = \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^7$$

Nhưng 3 lần xuất hiện mặt lục là tùy ý trong 10 lần gieo chứ không nhất thiết là lần thứ nhất, lần thứ tư và lần thứ chín như trên, tức là chỉ cần có 3 vị trí trong 10 vị trí là A. Ta có $C_{10}^3 = 120$ cách lấy ra 3 vị trí trong 10 vị trí. Vậy xác suất để có 3 lần xuất hiện mặt lục trong 10 lần gieo con xúc xắc sẽ là:

$$C_{10}^3 p^3 (1-p)^7 = C_{10}^3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^7$$

Với cách lý luận như vậy ta nhận được công thức (I.4).

I.5.3. Số có khả năng nhất

Khi thực hiện n phép thử Bernoulli, biến cố A có thể xảy ra từ 0 lần đến n lần. Theo (I.4) ta sẽ tính được (n + 1) xác suất tương ứng. Các xác suất này không như nhau, do đó sẽ có xác suất lớn nhất.

Số nguyên m_0 ($0 \le m_0 \le n$) được gọi là số có khả năng nhất nếu xác suất tương ứng với nó là lớn nhất:

$$P_n(m_0; p) = \max_{0 \le m \le n} P_n(m; p).$$

Như vậy, để tìm m_0, về trực giác, chúng ta phải tính (n + 1) xác suất theo công thức (I.4), từ đó rút ra giá trị lớn nhất. Công việc này sẽ khá nặng nề nếu n lớn, chẳng hạn n = 50; 100; 1000. Do đó, người ta đã tìm ra quy tắc để tìm số có khả năng nhất m_0 như sau:

Quy tắc tìm số m_0:

– Nếu (np + p – 1) là số nguyên thì m_0 = (np + p – 1) và (np + p).

– Nếu (np + p – 1) là số thập phân thì m_0 là số nguyên bé nhất nhưng lớn hơn số thập phân đó; m_0 = [np + p – 1] + 1 ([x] = phần nguyên của x).

Ví dụ I.6: Giả sử tỷ lệ người dân tham gia giao thông ở thành phố Hà Nội có hiểu biết cơ bản về luật giao thông là 80%. Nếu chọn ngẫu nhiên 20 người đang tham gia giao thông trên đường. Hãy tính xác suất của các tình huống sau:

a) Có 15 người hiểu biết luật giao thông.

b) Có 8 người không hiểu biết về luật giao thông.

c) Số người không hiểu biết về luật giao thông có khả năng nhất.

d) Trong một tình huống có 12 người đang bị cảnh sát giao thông xử lý vì vi phạm luật. Hãy đoán xem có bao nhiêu người hiểu biết luật giao thông nhưng cố tình vi phạm, bao nhiêu người vi phạm do không hiểu luật.

Giải:

Trước hết, cần xác định tình huống chọn ngẫu nhiên 20 người đang tham gia giao thông trên đường là chọn như thế nào? Đó là chọn từng người một và ta sẽ có 20 phép thử Bernoulli (xem nhận xét ở dưới), với A = {Chọn được người hiểu biết luật giao thông} và p = P(A) = 0,80.

a) $P_{20}(15; 0,8) = C_{20}^{15}.0,8^{15}.0,2^{5}$

b) $P_{20}(8; 0,20) = C_{20}^{8}.0,2^{8}.0,8^{12}$

(Có 8 người không hiểu luật tức là 8 lần xảy ra $\bar{A}$ với $P(\bar{A}) = 0,20$ nên ta có $P_{20}(8; P(\bar{A})) = P_{20}(8; 0,20)$.

Cách khác tương đương là: Có 8 người không hiểu luật, tức là có 12 người hiểu luật, nên xác suất là:

$$P_{20}(12; 0,80) = C_{20}^{12}.0,8^{12}.0,2^{8} = C_{20}^{8}.0,8^{12}.0,2^{8}$$

c) Ta quan tâm số người không hiểu biết luật, tức là số lần xảy ra $\bar{A}$ với xác suất 0,2; theo công thức $np + p - 1 = 20.0,2 + 0,2 - 1 = 3,2 \rightarrow m_0 = 4$.

Vậy, có 4 người không hiểu biết về luật trong số 20 người là con số có khả năng nhất.

d) Tình huống 12 người đang bị cảnh sát xử lý cũng coi như 12 phép thử Bernoulli với A và p như trên.

Để dự đoán, tất nhiên ta phải chọn tình huống xảy ra với xác suất cao nhất (để khả năng đúng là lớn nhất). Do đó, với người hiểu biết về luật ta có:

$np + p - 1 = 12.0,8 + 0,8 - 1 = 9,4 \rightarrow m_0 = 10$. Có 10 người hiểu biết về luật nhưng cố tình vi phạm.

Còn với người không hiểu luật ta có:

$np + p - 1 = 12 . 0,2 + 0,2 - 1 = 1.6 \rightarrow m_0 = 2$.

Có 2 người không hiểu luật nên vi phạm.

Nhận xét: Khi cho một tỷ lệ P(A) nào đó mà không cho biết số phần tử của tập đang xét thì phải hiểu là: Trong tình huống đó khả năng xảy ra A là như nhau trong các lần chọn, dù lấy lần thứ nhất hay lấy lần thứ n, dù có hoàn lại hay không hoàn lại (sự khác nhau giữa chúng coi như bỏ qua). Nếu lấy ra k phần tử từ tập đang xét với tình huống như thế thì phải hiểu là lấy từng phần tử một và lấy k lần (không thể hiểu là lấy cùng lúc được). Còn nếu lấy ra k phần tử từ một tập gồm n phần tử, nếu không nói gì thêm, thì lại phải hiểu là lấy một lần (lấy cùng lúc, lấy theo cách tổ hợp).

Bạn đọc phải nắm rõ điều này để đỡ lúng túng khi phân tích, xử lý bài toán.

I.6. BIẾN NGẪU NHIÊN

I.6.1. Định nghĩa

Một biến (hay một đại lượng) nhận các giá trị của nó với xác suất tương ứng nào đó được gọi là biến ngẫu nhiên, ký hiệu là X, Y, Z,...

Biến không phải là khái niệm xa lạ; bạn đọc đã biết đến các biến tất định, tức là loại biến chỉ nhận với xác suất 1 (trường hợp nó nhận) và xác suất 0 (trường hợp nó không nhận). Để hiểu kỹ hơn, bạn đọc hãy tự lý giải xem biến ngẫu nhiên rộng hơn biến tất định ở điểm nào).

Căn cứ vào tập giá trị mà biến ngẫu nhiên nhận, người ta phân chia biến ngẫu nhiên thành hai loại chính: biến ngẫu nhiên rời rạc và biến ngẫu nhiên liên tục.

I.6.2. Biến ngẫu nhiên rời rạc

Nếu các giá trị mà biến ngẫu nhiên nhận cách xa nhau một khoảng nào đó thì biến ngẫu nhiên được gọi là rời rạc.

Như vậy, để xác định biến ngẫu nhiên rời rạc chúng ta phải chỉ ra các giá trị nó nhận và xác suất nhận giá trị đó tương ứng. Một bảng với hai thông tin như vậy được gọi là bảng phân phối xác suất.

X	x_1	$x_2 \ldots x_i \ldots x_n$
$P(X = x_i)$	p_1	$p_2 \ldots p_i \ldots p_n$

Ta có:

$$\sum_i p_i = 1$$

hoặc có thể viết ngắn gọn (nếu chúng có quy luật):

$P(X = x_i) = p_i$; i = 1, 2, 3, ...

Ví dụ I.7: Gieo 3 đồng tiền cân đối, đồng chất. Gọi X là số mặt sấp xuất hiện. Hãy lập bảng phân phối xác suất của X.

Giải:

Dễ thấy X nhận 4 giá trị là: 0, 1, 2, 3.

Để tính 4 xác suất tương ứng, có thể dùng phương pháp cổ điển hoặc dùng xác suất Bernoulli.

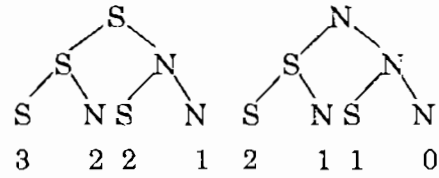
Theo phương pháp cổ điển, ta có số biến cố sơ cấp là 2.2.2 = 8.

8 biến cố này có thể mô tả như sau:

Đồng tiền thứ I:

Đồng tiền thứ II:

Đồng tiền thứ III:



Ta có bảng phân phối xác suất số mặt sấp xuất hiện như sau:

X	0	1	2	3
P(X = m)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Mặt khác, ta có 3 phép thử Berrnoulli với p = P (xuất hiện mặt sấp) = $\frac{1}{2}$ nên:

$$P(X = m) = P_3\left(m;\ \frac{1}{2}\right) = C_3^m\left(\frac{1}{2}\right)^3 ;\ m = 0, 1, 2, 3$$

Ví dụ I.8: Trong lô hàng 10 chiếc máy tính mới có 3 chiếc bị lỗi, lấy ngẫu nhiên ra 4 chiếc. Gọi Y là số máy tính bị lỗi trong 4 chiếc lấy ra. Hãy:

a) Lập bảng phân phối xác suất của Y.

b) Khi lấy ra 4 chiếc thì có mấy chiếc bị lỗi là có khả năng xảy ra cao nhất.

c) Tìm xác suất khi lấy ra 4 chiếc sẽ bị ít nhất 1 chiếc lỗi.

d) Nếu người mua lấy ngẫu nhiên ra 3 chiếc để kiểm tra, thấy không có chiếc nào bị lỗi thì sẽ chấp nhận cả lô hàng. Tìm xác suất người mua chấp nhận lô hàng. Bác bỏ lô hàng.

Giải:

Phép thử là lấy ra 4 máy tính trong lô 10 chiếc. Do đó, phép thử thực hiện một lần (lấy theo nghĩa tổ hợp). Số biến cố sơ cấp là $C_{10}^4 = 210$.

a)

Y	0	1	2	3
P(Y = m)	$\frac{C_3^0 C_7^4}{C_{10}^4}$	$\frac{C_3^1 C_7^3}{C_{10}^4}$	$\frac{C_3^2 C_7^2}{C_{10}^4}$	$\frac{C_3^3 C_7^1}{C_{10}^4}$
	$\frac{35}{210}$	$\frac{105}{210}$	$\frac{63}{210}$	$\frac{7}{210}$
	0,167	0,50	0,30	0,033

b) Theo bảng phân phối xác suất trên thì P(Y = 1) = 0,50 là cao nhất, cho nên trong 4 máy tính lấy ra bị 1 chiếc lỗi là tình huống xảy ra cao nhất.

c) $P(Y \geq 1) = 1 - P(Y = 0) = 1 - 0,167 = 0,833$

d) P(người mua chấp nhận lô hàng) =

= P(trong 3 máy tính lấy ra không có chiếc nào bị lỗi)

$$= \frac{C_3^0 C_7^3}{C_{10}^3} = \frac{35}{120} = 0,2917$$

P(người mua bác bỏ lô hàng) =

= P(có ít nhất 1 máy tính bị lỗi trong 3 chiếc lấy ra)

= 1 – 0,2917 = 0,7083.

Ví dụ I.9: Theo điều tra xã hội ở nước Anh, có 70% các ông chồng chưa hề làm công việc giặt là trong gia đình. Một phóng viên, tranh thủ lúc thời gian chờ lên tàu điện ngầm của hành khách, đã phỏng vấn một số nam hành khách. Anh ta dự định phỏng vấn tối đa 5 người, nhưng nếu gặp được nam giới đã từng tham gia giặt là giúp vợ thì thôi không phỏng vấn tiếp nữa. Gọi Z là số nam giới đã được phỏng vấn. Hãy mô tả quy luật phân phối của Z.

Giải:

Dễ thấy Z nhận các giá trị: 1, 2, 3, 4, 5.

Z	1	2	3	4	5
P(Z = k)	0,3	$0,7.0,3$	$0,7^2.0,3$	$0,7^3.0,3$	$0,7^4$

Z = 1, tức là phỏng vấn người đầu tiên, gặp ngay người đó đã từng giặt là giúp vợ.

Z = 3, tức là 2 người đầu chưa bao giờ giặt là, người thứ 3 đã từng làm.

Z = 5, tức là 4 người đầu chưa bao giờ giặt là, còn người thứ 5 có thể đã làm, có thể chưa làm. Dù tình huống nào cũng thôi, không phỏng vấn nữa, nên:

$$P(Z = 5) = 0,7^4 . 0,3 + 0,7^4 . 0,7 = 0,7^4$$

hoặc $P(Z = 5) = 1 - (P(Z = 1) + P(Z = 2) + P(Z = 3) + P(Z = 4))$

$$= 1 - 0,3(1 + 0,7 + 0,7^2 + 0,7^3)$$

$$= 1 - 0,7599 = 0,2401 = 0,7^4$$

Ví dụ I.10: Cho 2 biến ngẫu nhiên độc lập X và Y với các bảng phân phối như sau:

X	0	1	2
$P(X = x_i)$	0,2	0,5	0,3

Y	–1	1
$P(Y = y_i)$	0,4	0,6

Hãy lập bảng phân phối xác suất của biến ngẫu nhiên 2X, X^2 và X + Y.

Giải:

Trước hết, thế nào là sự độc lập của hai biến ngẫu nhiên?

Hai biến ngẫu nhiên gọi là độc lập với nhau nếu mọi biến cố liên quan đến biến này độc lập với biến cố bất kỳ liên quan đến biến kia.

$Z = 2X$	0	2	4
$P(Z = z_i)$	0,2	0,5	0.3

Hiển nhiên ta có: $P(kX = kx_i) = P(X = x_i)$ với $k \neq 0$

$Z = X^2$	0	1	4
$P(Z = z_i)$	0,2	0,5	0,3

$Z = X+Y$	−1	0	1	2	3
$P(Z = z_i)$	0.08	0,20	0,24	0,30	0,18

Vì $(Z = -1) = (X = 0)(Y = -1) \Rightarrow P(Z = -1) = 0.2.0,4 = 0,08$

$(Z = 1) = (X = 0)(Y = 1) \cup (X = 2)(Y = -1)$

Do tính xung khắc của 2 biến cố tích, do tính độc lập của X và Y nên:

$$P(Z = 1) = P((X = 0)(Y = 1)) + P((X = 2)(Y = -1))$$
$$= 0,2.0,6 + 0,3.0,4 = 0,24$$

I.6.3. Biến ngẫu nhiên liên tục

Nếu các giá trị mà biến ngẫu nhiên nhận lấp đầy một khoảng nào đó (nhận mọi giá trị trong khoảng đó) thì biến ngẫu nhiên được gọi là liên tục.

Để xác định biến ngẫu nhiên liên tục người ta dùng khái niệm hàm mật độ.

Hàm số p(x) được gọi là hàm mật độ của biến ngẫu nhiên nào đó nếu thỏa mãn:

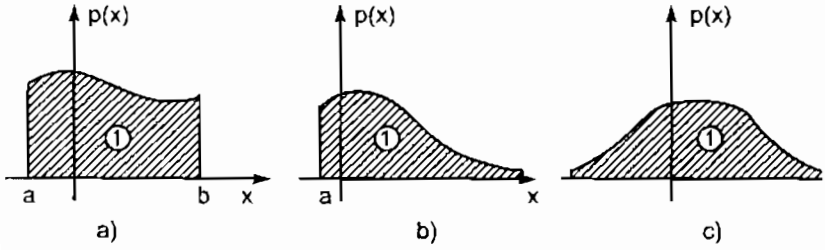
a) $p(x) \geq 0$

b) $\int_{-\infty}^{+\infty} p(x)dx = 1$

Thực chất hai điều kiện trên có nghĩa là phần diện tích giới hạn bởi đường cong mật độ p(x) (không nằm dưới trục hoành) và trục hoành là bằng 1.

Do đó nếu p(x) xác định ở vô hạn (+ ∞; – ∞ hoặc cả hai) thì đồ thị p(x) phải tiệm cận với trục hoành (để có phần diện tích hữu hạn).

Nhìn chung, đồ thị của hàm mật độ p(x) sẽ có các dạng sau:



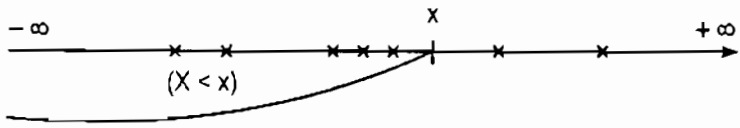
(a) Biến ngẫu nhiên xác định trên [a, b]
(b) Biến ngẫu nhiên xác định trên [a, + ∞)
(c) Biến ngẫu nhiên xác định trên (– ∞, + ∞)

I.7. HÀM PHÂN PHỐI

I.7.1. Định nghĩa

Hàm số F(x) = P(X < x) được gọi là hàm phân phối của biến ngẫu nhiên X. Trong đó, x là biến số thực của hàm F; x nhận giá trị – ∞ < x < + ∞.

Giá trị của hàm phân phối tại điểm x chính là xác suất để biến ngẫu nhiên nhận giá trị bên trái x (hình I.2).



Hình I.2. Giải thích bằng hình học về hàm phân phối
Các điểm (x) là các giá trị x_i mà biến ngẫu nhiên rời rạc X nhận.

I.7.2. Tính chất

– Miền xác định của hàm phân phối: ∀x ∈ (– ∞, + ∞)

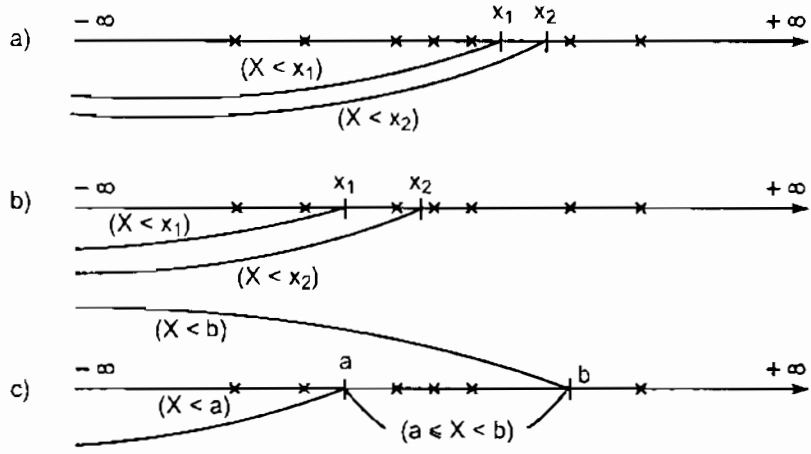
– Miền giá trị của hàm phân phối:

$0 \le F(x) \le 1, \forall x \in (-\infty, +\infty)$

$F(-\infty) = 0; F(+\infty) = 1$

– Hàm phân phối là hàm không giảm: nếu $x_1 < x_2$ thì $F(x_1) \le F(x_2)$

Ta có $P(a \le X < b) = F(b) - F(a)$



Hình I.3. (a) $x_1 < x_2$ nhưng $F(x_1) = F(x_2)$;
(b) $x_1 < x_2$ nhưng $F(x_1) < F(x_2)$; (c) $P(a \le X < b)$

Ta có:

$$F(x) = P(X < x) = \begin{cases} \sum\limits_{i: x_i < x} p_i & \text{nếu X rời rạc } P(X = x_i) = p_i \\ \int\limits_{-\infty}^{x} p(t)dt & \text{nếu X liên tục với mật độ } p(t) \end{cases} \quad (I.5)$$

Tổng trên được lấy với các chỉ số i mà $x_i < x$.

Nếu trong bảng phân phối xác suất, các giá trị x_i được xếp theo thứ tự tăng dần thì để tìm giá trị F(x) ta chỉ việc cộng dồn các giá trị p_i từ trái qua phải. Vì thế, F(x) còn được gọi là hàm phân phối tích lũy.

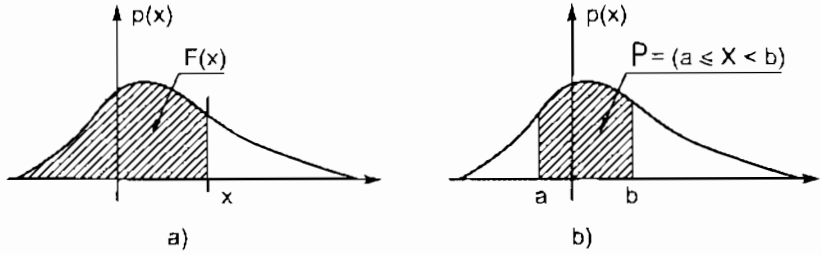
Từ (I.5) ta nhận được:

$$P(a \le X < b) = F(b) - F(a) = \begin{cases} \sum_{i:a \le x_i < b} p_i & \text{nếu } P(X = x_i) = p_i \\ \int_a^b p(t)dt & \text{nếu } p(t) \text{ là hàm mật độ} \end{cases} \quad (I.6)$$

Đối với biến ngẫu nhiên liên tục ta có:

$$P(a < X < b) = P(a < X < b)$$
$$= P(a \le X \le b)$$
$$= P(a < X \le b),$$

nghĩa là đối với biến liên tục, việc kín hay hở ở hai đầu mút khi tính xác suất đều cho kết quả như nhau.



Hình I.4. Minh họa hình học hàm phân phối (a) và xác suất $P(a \le X < b)$ đối với biến ngẫu nhiên liên tục (b).

Ví dụ I.11: Đối với mỗi ví dụ I.7, I.8, I.9 ở trên hãy:

a) Viết biểu thức hàm phân phối của các biến ngẫu nhiên X, Y, Z tương ứng.

b) Tính $P(0 \le X < 2,1)$; $P(1 < Y \le 3,6)$; $P(2 < Z \le 4,8)$

Giải:

Đối với biến X (Ví dụ I.7)

$$F(x) = \begin{cases} 0 & \text{nếu } x \le 0 \\ \frac{1}{8} & \text{nếu } 0 < x \le 1 \\ \frac{4}{8} & \text{nếu } 1 < x \le 2 \\ \frac{7}{8} & \text{nếu } 2 < x \le 3 \\ 1 & \text{nếu } 3 < x \end{cases}$$

$$P(0 \le X < 2,1) = F(2,1) - F(0)$$

$$= \frac{7}{8} - 0 = 0,875$$

hoặc

$$P(0 \le x < 2,1) = P(X = 0) + P(X = 1) + P(X = 2)$$

$$= \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}$$

$$= 0,875$$

Đối với biến Y (Ví dụ I.8):

$$F(x) = \begin{cases} 0 & \text{nếu } x \le 0 \\ 0,167 & \text{nếu } 0 < x \le 1 \\ 0,667 & \text{nếu } 1 < x \le 2 \\ 0,967 & \text{nếu } 2 < x \le 3 \\ 1,0 & \text{nếu } 3 < x \end{cases}$$

$$P(1 < Y \le 3,6) = P(Y = 2) + P(Y = 3)$$

$$= 0,30 + 0,033$$

$$= 0,333$$

Đối với biến Z (Ví dụ I.9):

$$F(x) = \begin{cases} 0 & \text{nếu } x \le 1 \\ 0,3 & \text{nếu } 1 < x \le 2 \\ 0,51 & \text{nếu } 2 < x \le 3 \\ 0,657 & \text{nếu } 3 < x \le 4 \\ 0,7599 & \text{nếu } 4 < x \le 5 \\ 1,0 & \text{nếu } 5 < x \end{cases}$$

$$P(2 < Z \le 4,8) = P(Z = 3) + P(Z = 4)$$
$$= 0,7^2.0,3 + 0,7^3.0,3 = 0,2499$$

Hoặc

$$P(2 < Z \le 4,8) = P(2 \le Z < 4,8) - P(Z = 2) + P(Z = 4,8)$$
$$= F(4,8) - F(2) - P(Z = 2) + P(Z = 4,8)$$
$$= 0,7599 - 0,30 - 0,7.0,3 + 0$$
$$= 0,2499$$

I.8. CÁC SỐ ĐẶC TRƯNG CỦA BIẾN NGẪU NHIÊN

I.8.1. Kỳ vọng (giá trị trung bình)

Định nghĩa:

Kỳ vọng của biến ngẫu nhiên X là một số, ký hiệu EX, được xác định như sau:

$$EX = \begin{cases} \sum_i x_i p_i & \text{nếu } P(X = x_i) = p_i \\ \int_{-\infty}^{+\infty} xp(x)dx & \text{nếu } p(x) \text{ là hàm mật độ} \end{cases}$$

Ý nghĩa:

Kỳ vọng của biến ngẫu nhiên là giá trị trung bình mà biến ngẫu nhiên nhận. Như vậy, kỳ vọng thực ra lại là một

khái niệm quen thuộc. Đó là giá trị trung bình. (E là viết tắt của từ expectation).

Bạn đã dùng kỳ vọng chưa và dùng từ khi nào? Những tình huống nào người ta dùng kỳ vọng? Bạn đọc hãy xem thêm phần phụ lục I ở cuối sách để xem suy nghĩ của bạn và tác giả có gần nhau không.

Có thể nói kỳ vọng là giá trị trung bình cộng hợp lý và khách quan nhất. Để thấy rõ ý nghĩa của kỳ vọng ta xét tình huống sau: Một gia đình chi tiêu trong 1 tháng ở hai mức: 5 triệu hoặc 4 triệu đồng, trong đó có 11 tháng ở mức 5 triệu, chỉ có 1 tháng ở mức 4 triệu. Hỏi trung bình 1 tháng gia đình này tiêu hết bao nhiêu tiền? Như vậy có hai mức chi tiêu (hai giá trị) cho nên trung bình số học (một cách đơn giản) sẽ là $\frac{1}{2}(5 + 4) = 4{,}5$ triệu đồng. Nhưng trung bình có trọng lượng sẽ là $5.\frac{11}{12} + 4.\frac{1}{12} = 4{,}917$ triệu đồng. Rõ ràng, trung bình có trọng lượng (cũng chính là cách tính của kỳ vọng toán) phản ánh chính xác hơn, hợp lý hơn trung bình số học.

Tính chất:

a) $EC = C$, $C = \text{const}$

b) $ECX = CEX$

c) $E(X \pm Y) = EX \pm EY$

d) Nếu X, Y độc lập thì $E(XY) = EX.EY$

e) $EX^2 = \begin{cases} \sum\limits_i x_i^2 p_i & \text{nếu } P(X = x_i) = p_i \\ \int\limits_{-\infty}^{+\infty} x^2 p(x)dx & \text{nếu X liên tục với mật độ } p(x) \end{cases}$

Như vậy, EX là một giá trị thực, có thể dương, có thể bằng 0 và cũng có thể âm.

Biến ngẫu nhiên X có thể nhận các giá trị là nguyên, nhưng giá trị trung bình EX có thể là số thập phân. Ngược lại, nếu EX là một giá trị thập phân thì không có nghĩa là X nhận giá trị thập phân. Chẳng hạn, điểm thi trung bình các môn học trong học kỳ của một sinh viên là 7,18 nhưng điểm thi của từng môn lại là các giá trị nguyên. Hoặc theo ví dụ I.7, $EX = \frac{12}{8} = 1,5$, tức là số mặt sấp xuất hiện trung bình là 1,5 nhưng số mặt sấp xuất hiện ở mỗi lần gieo sẽ phải là số nguyên: hoặc 0, hoặc 1, hoặc 2, hoặc 3.

I.8.2. Phương sai

Định nghĩa:

Phương sai của biến ngẫu nhiên X là một số không âm, ký hiệu DX, được xác định như sau:

$$DX = E(X - EX)^2$$
$$= EX^2 - (EX)^2$$

(chữ D là viết tắt của từ dispersion)

Ý nghĩa (Để tránh trùng lặp, mời bạn đọc xem ở phần Phụ lục I).

Tính chất:

a) $DC = 0$; $C = const$

b) $DCX = C^2DX$

c) $D(X \pm Y) = DX + DY$ nếu X, Y độc lập

d) $D(aX + b) = a^2DX$

Ta ký hiệu $\sigma = \sqrt{DX}$ và gọi là độ lệch tiêu chuẩn.

I.8.3. Mode

Mode của biến ngẫu nhiên X là một giá trị của biến ngẫu nhiên X, ký hiệu ModX, mà tại đó biến ngẫu nhiên X nhận với xác suất lớn nhất (nếu X rời rạc) hoặc tại đó hàm mật độ đạt cực đại (nếu X liên tục).

I.8.4. Median và phân vị

– Median (trung vị) của biến ngẫu nhiên X là một số, ký hiệu MedX, được xác định như sau:

$$\begin{cases} P(X < MedX) \le \dfrac{1}{2} \\ P(X \le MedX) \ge \dfrac{1}{2} \end{cases}$$

Nếu X là biến ngẫu nhiên liên tục thì Median xác định từ: $F(MedX) = \dfrac{1}{2}$; nghĩa là MedX chia miền giá trị của biến ngẫu nhiên X thành hai nửa có xác suất bằng nhau $\left(\dfrac{1}{2}\right)$.

– Phân vị cấp p: x_p được gọi là phân vị cấp p của biến ngẫu nhiên X nếu:

$$\begin{cases} P(X < x_p) \le p \\ P(X \le x_p) \ge p \end{cases}$$

Nếu X là biến ngẫu nhiên liên tục thì ta có $F(x_p) = p$.

Người ta xác định các tứ phân vị: $x_{\frac{1}{4}}$, $x_{\frac{2}{4}}$, $x_{\frac{3}{4}}$ $(x_{\frac{2}{4}} = x_{\frac{1}{2}} = MedX)$.

Các thập phân vị: $x_{\frac{1}{10}}$, $x_{\frac{2}{10}}$, ..., $x_{\frac{9}{10}}$.

Các bách phân vị: $x_{\frac{k}{100}}$, k = 1, 2..., 99.

– Khoảng tứ phân vị: $\left(x_{\frac{1}{4}}, x_{\frac{3}{4}}\right)$.

Rõ ràng ta có $P(X \in (x_{\frac{1}{4}}, x_{\frac{3}{4}})) = 50\%$ nếu X là biến ngẫu nhiên liên tục, nên người ta cũng dùng khoảng tứ phân vị để đo mức độ tập trung, phân tán của biến ngẫu nhiên.

Ví dụ I.12: Trở lại ví dụ I.10: Hãy tính EZ, DZ, σ, ModZ, trong đó Z = X + Y.

Giải:

$EZ = (-1).0,08 + 0 + 1.0,24 + 2.0,30 + 3.0,18 = 1,30$

$EZ^2 = (-1)^2.0,08 + 0 + 1^2.0,24 + 2^2.0,30 + 3^2.01,8 = 3,14$

$DZ = 3,14 - 1,3^2 = 1,45$

$\sigma = \sqrt{1,45} = 1,204$

ModZ = 2

I.9. MỘT VÀI PHÂN PHỐI CẦN DÙNG

Trong xác suất thống kê người ta hay ký hiệu $X = F(x)$, nghĩa là biến ngẫu nhiên X có hàm phân phối là F(x).

Ở mức độ giáo trình Lý thuyết Xác suất cơ sở (30 tiết; 45 tiết trở lên) thì các phân phối xác suất thông dụng sau cần phải đề cập đến: phân phối nhị thức, phân phối siêu bội, phân phối hình học, phân phối Poisson, phân phối mũ, phân phối chuẩn, phân phối đều, phân phối khi bình phương, phân phối Student, phân phối F của Fisher. Nhưng ở mức độ của giáo trình này tác giả chỉ giới thiệu 2 phân phối rời rạc đơn

giản hay gặp trong thực tế là: phân phối nhị thức, phân phối siêu bội; đồng thời đề cập 3 phân phối cần dùng đến trong phần thống kê ứng dụng ở chương sau là: phân phối chuẩn, phân phối khi bình phương và phân phối Student. Phân phối chuẩn cũng là phân phối có nhiều ứng dụng trong thực tế.

I.9.1. Phân phối nhị thức B(n; p)

Xét n phép thử Bernoulli với biến cố A có P(A) = p.

Gọi X là biến ngẫu nhiên chỉ số lần xảy ra biến cố A trong n phép thử Bernoulli nói trên. Phân phối của biến ngẫu nhiên X được gọi là phân phối nhị thức, ký hiệu B(n; p) (B là viết tắt của từ binomial).

$$P(X = m) = C_n^m p^m (1-p)^{n-m}; \quad m = \overline{0;n}$$

Ta có: $EX = np$; $DX = np(1-p)$;

$ModX = m_0$ (xem I.5.3 trang 24).

Để chứng minh các kết quả trên chúng ta tính toán theo định nghĩa. Nhưng dưới đây giới thiệu một cách tính khá đặc biệt.

Ta xây dựng n biến ngẫu nhiên ứng với n phép thử Bernoulli như sau:

$$X_i = \begin{cases} 1 \text{ nếu ở phép thử thứ i biến cố A xảy ra} \\ 0 \text{ nếu ở phép thử thứ i biến cố } \bar{A} \text{ xảy ra} \end{cases}$$

$$X_i = \begin{cases} 1 \text{ với xác suất p} \\ 0 \text{ với xác suất } 1-p \end{cases}$$

Rõ ràng: $X_1, X_2, \ldots, X_n$ độc lập, cùng phân phối như nhau và $X = \sum_{i=1}^{n} X_i$.

Ta có: $EX_i = 1.p + 0.(1 - p) = p$

$EX_i^2 = 1^2.\ p + 0 = p$

$DX_i = p - p^2 = p(1 - p)$

$$EX = \sum_{i=1}^{n} EX_i = n.p$$

$$DX = \sum_{i=1}^{n} DX_i = np(1 - p)$$

Còn theo định nghĩa thì ModX chính là số có khả năng nhất m_0.

Ví dụ I.13: Theo một điều tra về xã hội học cho thấy tỷ lệ sinh viên học tập không đúng với ngành nghề mà họ yêu thích là 34%. Một lớp gồm 60 sinh viên. Gọi X là số sinh viên không theo đúng ngành nghề yêu thích trong 60 sinh viên này. Hãy:

a) Mô tả quy luật phân phối của X.

b) Về trung bình thì trong 60 sinh viên sẽ có bao nhiêu sinh viên không thích ngành đang học?

c) Có bao nhiêu sinh viên không thích ngành đang học là có khả năng nhất?

Giải:

a) Lớp gồm 60 sinh viên được coi như ta đã chọn ngẫu nhiên ra 60 sinh viên từ tập tất cả sinh viên, với tỷ lệ không thích ngành đang học là 34%. Do đó, ta có 60 lần chọn với p = 0,34; nghĩa là ta có 60 phép thử Bernoulli với p = 0,34 (xem nhận xét ở I.5, trang 26). Theo định nghĩa ta có:

$$X \approx B(60\ ;\ 0{,}34) \leftrightarrow P(X = m) = C_{60}^{m}\ .0{,}34^{m}\ .\ 0{,}66^{60-m}$$

$$m = \widetilde{0;60}$$

b) Về trung bình trong 60 sinh viên sẽ có EX = 60.0,34 = 20,4 sinh viên không thích ngành đang học.

c) $np + p - 1 = 60.0,34 + 0,34 - 1 = 19,74 \Rightarrow m_0 = 20$

Trong 60 sinh viên thì tình huống có 20 sinh viên không thích ngành đang học là có khả năng xảy ra cao nhất.

I.9.2. Phân phối siêu bội (siêu hình học)

Phân phối của biến ngẫu nhiên Y trong ví dụ I.8 chính là phân phối siêu bội. Giả sử ta có một tập gồm N phần tử, trong đó có M phần tử mang dấu hiệu A nào đó, còn lại (N – M) phần tử mang dấu hiệu $\overline{A}$. Lấy ngẫu nhiên ra n phần tử từ tập đã cho. Gọi X là số phần tử mang dấu hiệu A trong n phần tử được lấy ra. Phân phối của X được gọi là phân phối siêu bội.

$$P(X = m) = \frac{C_M^m \, C_{N-M}^{n-m}}{C_N^n} \; ; \; m \le \min(M; n)$$

Phép thử ở đây là lấy một lần, lấy cùng lúc, cho nên theo xác suất cổ điển chúng ta dễ dàng có được phân phối xác suất như trên. Việc tìm các số đặc trưng tính trực tiếp từ bảng phân phối xác suất sẽ dễ dàng hơn là nhớ và dùng công thức (xem [1]).

Ví dụ I.14: Một tổ gồm 12 người, trong đó có 5 nữ. Cần cử 4 người đi làm một việc. Để công bằng người ta đã thực hiện chọn ngẫu nhiên. Gọi X là số nữ trong 4 người được chọn. Hãy:

a) Lập bảng phân phối xác suất của X.

b) Viết biểu thức hàm phân phối của X.

c) Tính EX, DX, ModX, E(3X – 2), D(5 – 3X).

Giải:

a) Phép thử là lấy cùng lúc ra 4 người, tức là lấy theo nghĩa tổ hợp.

Ta có

X	0	1	2	3	4
P(X=m)	$\frac{C_5^0 C_7^4}{C_{12}^4}$	$\frac{C_5^1 C_7^3}{C_{12}^4}$	$\frac{C_5^2 C_7^2}{C_{12}^4}$	$\frac{C_5^3 C_7^1}{C_{12}^4}$	$\frac{C_5^4 C_7^0}{C_{12}^4}$
	$\frac{35}{495}$	$\frac{175}{495}$	$\frac{210}{495}$	$\frac{70}{495}$	$\frac{5}{495}$
	0,071	0,354	0,424	0,141	0,010

b)
$$F(x) = \begin{cases} 0 & \text{nếu } x \le 0 \\ 0{,}071 & \text{nếu } 0 < x \le 1 \\ 0{,}425 & \text{nếu } 1 < x \le 2 \\ 0{,}849 & \text{nếu } 2 < x \le 3 \\ 0{,}990 & \text{nếu } 3 < x \le 4 \\ 1{,}0 & \text{nếu } 4 < x \end{cases}$$

c) $EX = 0 + 1.0{,}354 + 2.0{,}424 + 3.0{,}141 + 4.0{,}010$

$= 1{,}665$

$EX^2 = 0 + 1.0{,}354 + 2^2.0{,}424 + 3^2.0{,}141 + 4^2.0{,}010$

$= 3{,}479$

$DX = 3{,}479 - 1{,}665^2 = 0{,}706775$

$ModX = 2$

$E(3X - 2) = 3EX - 2 = 3.1{,}665 - 2 = 2{,}995$

$D(5 - 3X) = 3^2\, DX = 9.0{,}706775 = 6{,}360975$

I.9.3. Phân phối chuẩn $N(\mu; \sigma^2)$

Đây là biến ngẫu nhiên liên tục, nhận giá trị trên khoảng $(-\infty, +\infty)$ với hàm mật độ có dạng:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad ; -\infty < x < +\infty$$

trong đó: $EX = ModX = \mu$; $DX = \sigma^2$.

Chữ N là viết tắt của từ Normal.

Để chứng minh hai kết quả trên chúng ta phải tính tích phân suy rộng, do đó ta bỏ qua và công nhận kết quả (xem [1]).

Xét phân phối chuẩn tiêu chuẩn N(0; 1):

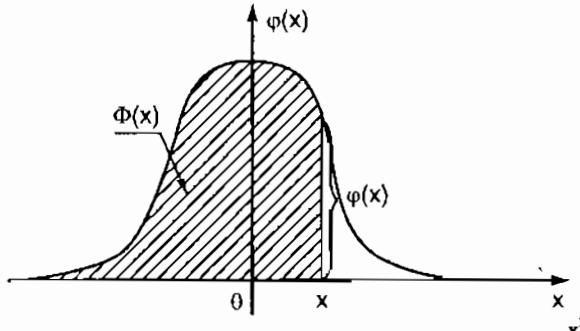
Hàm mật độ của nó sẽ là:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} . e^{-\frac{x^2}{2}}, -\infty < x < +\infty.$$

Theo (I.6) hàm phân phối tương ứng là:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$

Vì tích phân trên không có nguyên hàm cho nên để tính giá trị $\Phi(x)$ người ta phải dùng phương pháp tính xấp xỉ và phải lập trình để thực hiện tính toán trên máy tính điện tử. Các kết quả tính toán đã được lập thành bảng (Bảng I – Phụ lục II). Chúng ta chỉ việc tra bảng để dùng.



Hình I.5. Đường cong mật độ N(0,1): $\varphi(x) = \frac{1}{\sqrt{2\pi}} . e^{-\frac{x^2}{2}}$

Giá trị $\Phi(x)$ = diện tích phần gạch bên trái x.

Ta có kết quả sau:

Nếu $X \simeq N(\mu; \sigma^2)$ thì $\dfrac{X-\mu}{\sigma} \simeq N(0; 1)$

Do đó: $P\{a \le X \le b\} = P\left\{\dfrac{a-\mu}{\sigma} \le \dfrac{X-\mu}{\sigma} \le \dfrac{b-\mu}{\sigma}\right\}$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \qquad (I.7)$$

Và lưu ý là: $\Phi(-x) + \Phi(x) = 1$.

Ví dụ I.15: Gọi X là chỉ số thông minh - IQ (Intelligent Quota) của học sinh lứa tuổi 12 – 15. Giả sử $X \simeq N(85; 25)$.

a) Cho biết chỉ số IQ trung bình của học sinh là bao nhiêu?

b) Tính xác suất chọn được học sinh rất thông minh ($X \ge 90$).

c) Tính tỷ lệ học sinh có chỉ số IQ $\in (80 ; 95)$.

d) Gọi Y là số học sinh có IQ $\in (80 ; 95)$ trong lớp 50 học sinh. Hãy chỉ rõ phân phối xác suất của Y.

e) Trong một lớp gồm 50 học sinh thì trung bình có bao nhiêu em rất thông minh ($X \ge 90$)?

Con số trung bình tìm được có phải là số có khả năng xảy ra cao nhất hay không? Vì sao?

Giải:

a) Chỉ số IQ trung bình chính là EX và bằng 85.

b) $P(X \ge 90) = P(90 \le X < +\infty) = \Phi(+\infty) - \Phi\left(\dfrac{90-85}{5}\right)$

$$= 1 - \Phi(1) = 1 - 0{,}8413 = 0{,}1587$$

c) Tỷ lệ học sinh có IQ $\in (80; 95)$ chính là $P(80 < X < 95)$.

Theo (I.7) ta có:

$$P(80 < X < 95) = \Phi\left(\frac{95-85}{5}\right) - \Phi\left(\frac{80-85}{5}\right)$$

$$= \Phi(2) - \Phi(-1) = \Phi(2) - (1 - \Phi(1))$$

$$= 0{,}9773 + 0{,}8413 - 1 = 0{,}8186 \approx 82\%$$

d) Một lớp gồm 50 học sinh được chọn từ tập học sinh với tỷ lệ $p = P(80 < X < 95) \approx 82\%$ được xem như 50 phép thử Bernoulli với xác suất $p = 82\%$ (xem nhận xét ở I.5, trang 26). Do đó, Y có phân phối nhị thức B(50; 0,82), tức là:

$$P(Y = m) = C_{50}^{m}\ 0{,}82^{m}\ 0{,}18^{50-m}; \quad m = \overline{0;50}$$

e) Tương tự như câu d), ta lại có 50 phép thử Bernoulli với $p = 0{,}1587$ nên trung bình trong lớp sẽ có $np = 50.0{,}1587 = 7{,}935$ học sinh (hoặc gần 8 học sinh) rất thông minh.

Số trung bình trên không phải là số có khả năng xảy ra cao nhất, vì $np + p - 1 = 50.0{,}1587 + 0{,}1587 - 1 = 7{,}0937 \Rightarrow m_0 = 8$, tức là số trung bình là 7,935. Còn số có khả năng nhất là 8. (Có thể thấy ngay số trung bình 7,935 không phải là số có khả năng nhất vì số có khả năng nhất m_0 phải là số nguyên từ 0 đến 50, mà ở đây số trung bình lại là số thập phân).

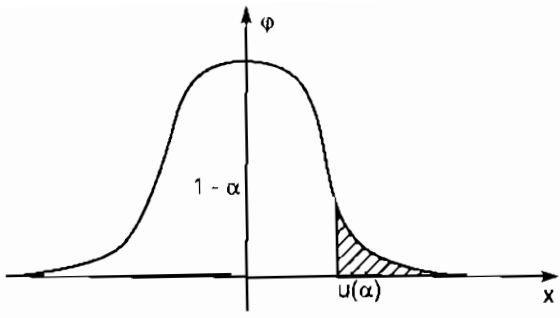
(Câu d) và e) có mức khó hơn đối với mức độ giáo trình này).

Ta ký hiệu $u(\alpha)$ là giá trị xác định từ phương trình:

$\Phi(u(\alpha)) = 1 - \alpha$

Nếu $X \simeq N(\mu\ ;\ \sigma^2)$ thì ta có:

$$P\left(\frac{X-\mu}{\sigma} < u(\alpha)\right) = 1 - \alpha$$



Hình I.6. Giá trị u(α).
Phần gạch chéo có diện tích bằng α

Hoặc: $P(X < \mu + \sigma\, u(\alpha)) = 1 - \alpha$ (I.8)

Hoặc: $$P\left(\frac{X-\mu}{\sigma} \geq u(\alpha)\right) = \alpha$$

$$P(X \geq \mu + \sigma\, u(\alpha)) = \alpha \quad \text{(I.9)}$$

Hoặc: $$P\left(\frac{|X-\mu|}{\sigma} < u\left(\frac{\alpha}{2}\right)\right) = 1 - \alpha$$

$$P\left(\mu - \sigma u\left(\frac{\alpha}{2}\right) < X < \mu + \sigma u\left(\frac{\alpha}{2}\right)\right) = 1 - \alpha \quad \text{(I.10)}$$

Hoặc: $$P\left(\frac{|X-\mu|}{\sigma} \geq u\left(\frac{\alpha}{2}\right)\right) = \alpha$$

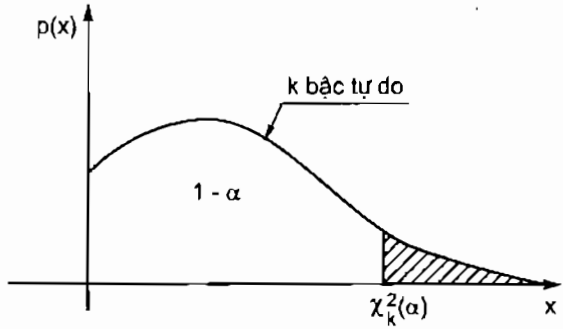
Các giá trị u(α) được tìm bằng cách tra bảng I (Phụ lục II).

Từ (I.10) chúng ta nhận được quy tắc kσ (xem Phụ lục I).

I.9.4. Phân phối khi bình phương χ^2_k

Đây là phân phối liên tục, xác định trên [0, +∞), hàm

mật độ phụ thuộc vào tham số nguyên dương k (mà người ta thường nói là mật độ với bậc tự do k). Biểu thức hàm mật độ lại liên quan tới hàm Gamma (chúng ta chưa được làm quen), cho nên tác giả không thấy cần thiết phải đưa biểu thức hàm mật độ này (xem [1]). Nhưng chỉ với các thông tin trên, theo I.6.3 và hình vẽ I.1, chúng ta cũng có thể vẽ phác được đường cong mật độ này (hình I.7).



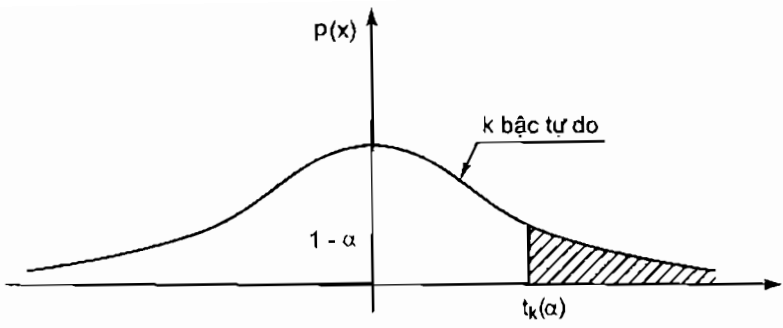
Hình I.7. Phác họa đường cong mật độ χ^2_k và giá trị $\chi^2_k(\alpha)$.
Phần gạch chéo có diện tích bằng α

Các giá trị $\chi^2_k(\alpha)$ được xác định như trong hình I.7, đã được tính sẵn và lập thành bảng (Bảng II – Phụ lục II).

I.9.5. Phân phối Student t_k

Đây là phân phối liên tục, xác định trên $(-\infty, +\infty)$, mật độ phụ thuộc tham số k (k nguyên, dương) hay mật độ với bậc tự do k, đường cong mật độ đối xứng qua trục tung. Tương tự như phân phối χ^2_k chúng ta vẽ phác được đường cong mật độ này (hình I.8).

Các giá trị $t_k(\alpha)$, được xác định như trong hình I.8, đã được tính sẵn và lập thành bảng (Bảng III – Phụ lục II).



Hình I.8. Phác họa đường cong mật độ phân phối Student và giá trị $t_k(\alpha)$. Phần gạch chéo có diện tích bằng α.

Từ hình I.7 và I.8 chúng ta cần hiểu ý nghĩa xác suất của các giá trị $\chi^2_k(\alpha)$ cũng như $t_k(\alpha)$. Chẳng hạn $t_{11}(0,05) = 1,80$ nghĩa là :

P (biến ngẫu nhiên $t_{11} > 1,80$) = 0,05

P (biến ngẫu nhiên $t_{11} < 1,80$) = 0,95

Ta có:

$$P\left\{|t_k| < t_k\left(\frac{\alpha}{2}\right)\right\} = P\left\{-t_k\left(\frac{\alpha}{2}\right) < t_k < t_k\left(\frac{\alpha}{2}\right)\right\} = 1-\alpha \ . \tag{I.11}$$

Hoặc $$P\left\{|t_k| \geq t_k\left(\frac{\alpha}{2}\right)\right\} = \alpha . \tag{I.12}$$

BÀI TẬP CHƯƠNG I

1. Đoàn đại biểu Quốc hội của tỉnh nọ gồm 10 người, trong đó có 3 đại biểu là người dân tộc ít người. Chọn ngẫu nhiên trong đoàn ra 3 người. Tìm khả năng xảy ra các tình huống sau:

a) Chọn được cả 3 đại biểu dân tộc ít người.

b) Chọn được 2 đại biểu dân tộc ít người.

c) Chọn được 1 đại biểu dân tộc ít người.

d) Không chọn được đại biểu nào thuộc dân tộc ít người.

e) Trước khi chọn hãy đoán xem sẽ chọn được bao nhiêu đại biểu dân tộc ít người?

2. Lấy ngẫu nhiên ra 6 con bài từ bộ tú lơ khơ 52 con. Tìm xác suất của các biến cố sau:

a) Lấy được 4 con màu đỏ.

b) Lấy được 1 con cơ, 2 con rô, 1 con pic.

c) Lấy được 1 con At, 3 con K và 2 con 9.

3. Hùng có một hộp kẹo, trong đó có 5 chiếc kẹo cốm, 6 chiếc kẹo sôcôla (chocolate), 4 chiếc kẹo hoa quả. Có hai bạn đến chơi, Hùng lấy kẹo ra mời bạn. Lần đầu Hùng lấy ngẫu nhiên ra 3 chiếc để mỗi người 1 chiếc. Tìm khả năng xảy ra các tình huống sau:

a) Lấy được cả ba loại kẹo.

b) Lấy được 2 kẹo cốm, 1 kẹo sôcôla.

c) Lấy được 3 chiếc sôcôla.

d) Lấy được 2 kẹo hoa quả, 1 kẹo cốm.

4. Gieo 2 con xúc xắc cân đối, đồng chất. Gọi X là tổng số chấm xuất hiện ở mặt trên của 2 con xúc xắc. Hỏi:

a) Số biến cố sơ cấp.

b) Tính $P(1 \le X < 5)$, $P(10 \le X < 15)$, $P(12 < X < 15)$.

5. Công đoàn trường tổ chức 3 hoạt động nhân ngày 8/3: cắm hoa, nấu ăn và khiêu vũ. Một bộ môn có 8 cô giáo đều ghi tên tham gia và mỗi người đăng ký một hoạt động một cách ngẫu nhiên (khả năng chọn 3 hoạt động là như nhau) và độc lập. Tìm xác suất xảy ra các tình huống sau:

a) Cả 8 người đều đăng ký vào khiêu vũ.

b) Cả 8 người cùng đăng ký vào một hoạt động.

c) Có 5 người cắm hoa, 2 người nấu ăn, 1 người khiêu vũ.

d) Hai cô giáo A và B cùng đăng ký vào một hoạt động.

6. Hai sinh viên Z và T thi môn Thống kê xã hội học. Khả năng đạt của mỗi người tương ứng là 0,85 và 0,95. Tìm xác suất xảy ra các tình huống sau:

a) Cả hai đều đạt.

b) Có ít nhất một người đạt.

c) Cả hai đều không đạt.

7. Thủ trưởng muốn triển khai một công việc ở hai đơn vị A và B. Biết rằng khả năng hoàn thành công việc đúng thời hạn của 2 đơn vị tương ứng là 95% và 80%. Tìm khả năng xảy ra các tình huống sau:

a) Cả hai đơn vị hoàn thành công việc đúng thời hạn.

b) Có ít nhất một đơn vị hoàn thành công việc đúng thời hạn.

c) Chỉ có đơn vị A hoàn thành công việc đúng thời hạn.

d) Có đúng một đơn vị hoàn thành công việc đúng thời hạn.

8. Theo thống kê của ngành Tòa án, tỷ lệ các vụ ly hôn do ngoại tình là 30%. Hãy tính xác suất để trong 100 vụ ly hôn được xử trong một năm qua ở thành phố Z có (chỉ yêu cầu viết được công thức tính):

a) 40 vụ ly hôn vì ngoại tình.

b) 50 vụ ly hôn vì lý do khác.

c) Có ít nhất một vụ ly hôn do ngoại tình.

d) Số vụ ly hôn do ngoại tình có khả năng nhất.

9. Theo con số thống kê công bố năm 1997, ở Việt Nam tỷ lệ các cặp vợ chồng vô sinh là 1/8, trong đó 40% trường hợp là do từ phía người chồng. Từ sổ đăng ký kết hôn ở một phường cách đây 10 năm cho thấy có 1240 cặp vợ chồng.

a) Hãy đoán xem có bao nhiêu cặp vô sinh? Viết công thức tính xác suất xảy ra trường hợp đó.

b) Trong số các cặp vô sinh dự đoán trên, có bao nhiêu cặp do chồng là có khả năng nhất?

c) Tính xác suất có không quá 2 cặp vô sinh do vợ.

10. Một người có 5 viên đạn, bắn từng viên vào bia cho đến khi có được 1 viên trúng tâm thì thôi không bắn nữa. Xác suất bắn trúng tâm của mỗi lần bắn là 0,70. Gọi X là số đạn người này đã bắn. Hãy:

a) Lập bảng phân phối xác suất của X.

b) Viết biểu thức hàm phân phối.

c) Tính EX, DX, ModX, $P(1 \le X \le 3)$, $P(2 < X \le 5,4)$.

11. Cho 2 biến ngẫu nhiên độc lập X, Y với bảng phân phối xác suất như sau:

X	−2	−1	0	2
$P(X = x_i)$	0,1	0,3	0,4	0,2

Y	−1	1
$P(Y = y_j)$	0,3	0,7

a) Lập bảng phân phối xác suất của 3X, X^2, X + Y.

b) Viết biểu thức hàm phân phối của X + Y.

c) Tính E(X + Y), D(X − Y), Mod(X + Y) và

$P(-2 < X + Y \le 2,1)$.

12. Tỷ lệ trẻ bị suy dinh dưỡng ở các tỉnh Tây Nguyên là 35,8%. Khảo sát ngẫu nhiên 50 trẻ ở vùng này. Gọi X là số trẻ suy dinh dưỡng (SDD) trong 50 trẻ nói trên. Hỏi:

a) X có phân phối gì?

b) Về trung bình trong 50 trẻ sẽ có bao nhiêu trẻ SDD?

c) Số trẻ SDD có khả năng xảy ra cao nhất là bao nhiêu?

d) Có khả năng xảy ra tình huống cả 50 trẻ được xét đều bị SDD hay không?

e) Tìm xác suất để có ít nhất một trẻ không bị SDD.

13. Theo thống kê, tỷ lệ trẻ em dưới 5 tuổi thừa cân năm 2004 là 1,7%. Trong một trường mầm non có 120 cháu. Hỏi:

a) Số trẻ thừa cân ở nhà trẻ này có phân phối gì?

b) Bao nhiêu trẻ thừa cân là tình huống xảy ra cao nhất?

c) Số tìm được trên có phải là số trẻ thừa cân trung bình trong 120 trẻ hay không?

d) Qua câu b) và c), liệu có xảy ra tình huống có trên 10 trẻ bị thừa cân trong số 120 trẻ hay không?

14. Trong một phòng có 12 người, trong đó có 4 người không thích xem bóng đá. Chọn ngẫu nhiên ra 5 người. Gọi X là số người không thích xem bóng đá trong 5 người chọn ra. Hãy:

a) Mô tả phân phối xác suất của X.

b) Viết biểu thức phân phối của X.

c) Tính EX, DX, ModX, $P(X \leq 2,3)$.

15. Gọi Y (cm) là chiều cao của các em lứa tuổi 13. Giả sử $Y \simeq N(140; 9)$.

a) Từ giả thiết đã cho chúng ta nhận được những thông tin gì về chiều cao của các em lứa tuổi 13?

b) Tỷ lệ các em lứa tuổi 13 có chiều cao thuộc khoảng (135; 150) là xác suất của biến cố nào? Tính tỷ lệ đó.

c) Gọi Z là số học sinh lớp 7 (tuổi 13) có chiều cao thuộc (135cm; 150cm) trong một lớp gần 50 em. Hãy:

i) Tìm phân phối xác suất của Z.

ii) Về trung bình, trong một lớp 50 em sẽ có bao nhiêu em cao từ 135cm đến 150cm?

iii) Hãy đoán xem trong một lớp 50 em sẽ có bao nhiêu em cao từ 135cm đến 150cm. Vì sao?

Chương II

THỐNG KÊ XÃ HỘI

II.1. GIỚI THIỆU BÀI TOÁN

Như trong ví dụ I.15 hoặc ví dụ 3 (Phụ lục I) chúng ta đã xét biến ngẫu nhiên chỉ số IQ, ký hiệu là X. Nếu ta biết biến ngẫu nhiên X có phân phối N(85; 25) thì theo các kết quả cơ bản của lý thuyết xác suất như đã giới thiệu ở chương I (hoặc ở Phụ lục I) chúng ta có thể tính được các tỷ lệ học sinh có chỉ số IQ thuộc vào một khoảng hoặc bán khoảng nào đó v.v.... Nhưng vấn đề đặt ra là: làm sao chúng ta biết biến ngẫu nhiên chỉ số IQ X có phân phối chuẩn với tham số 85 và 25? Lý thuyết xác suất không trả lời được câu hỏi này và lẩn tránh nó, trong bài toán xác suất nó được phát biểu dưới dạng giả thiết đã cho. Thống kê Toán học sẽ giải quyết và trả lời câu hỏi trên.

Có thể nói Thống kê xã hội thực chất là những kết quả đơn giản của Thống kê ứng dụng (một phần của Thống kê Toán học) được dùng trong nghiên cứu xã hội.

Giả sử ta cần nghiên cứu một đặc trưng xã hội nào đó, ta biết đó là một biến ngẫu nhiên. Theo kết quả của phần xác suất, khi đề cập đến biến ngẫu nhiên chúng ta biết rằng quy

luật phân phối xác suất của nó sẽ cho thông tin đầy đủ nhất về biến ngẫu nhiên này và chúng ta cố gắng tìm chúng. Sau quy luật phân phối xác suất thì các số đặc trưng của biến ngẫu nhiên sẽ cho ta một phần thông tin quan trọng. Ở mức độ của giáo trình Thống kê xã hội học này chúng ta không thể có câu trả lời về phân phối xác suất, mà chỉ có thể tìm câu trả lời cho các số đặc trưng, chính xác hơn là tìm câu trả lời cho hai đại lượng quen thuộc và được sử dụng rộng rãi là giá trị trung bình và tỷ lệ.

Tìm tỷ lệ và tính giá trị trung bình là hai bài toán quen thuộc mà chúng ta đã biết khi học ở Tiểu học và THCS. Đó là tình huống tính tỷ lệ, tính giá trị trung bình khi thông tin mà chúng ta có là đầy đủ. Giá trị tỷ lệ và trung bình mà chúng ta tính được là giá trị tỷ lệ thực, giá trị trung bình thực của nó. Còn trong giáo trình này, giá trị trung bình thực, ký hiệu là EX hoặc μ, tỷ lệ thực hay giá trị xác suất thực, ký hiệu $P(X \in A)$ hoặc p, là chưa biết và đang cần tìm câu trả lời. Theo bạn, liệu câu trả lời của chúng ta có giống như câu trả lời ở tình huống có thông tin đầy đủ?

Để có câu trả lời chúng ta phải thu thập thông tin về biến ngẫu nhiên X mà ta đang xét. Nhưng thông tin ta cố gắng thu thập được sẽ không phải là thông tin đầy đủ, mà chỉ là thông tin mẫu đại diện. Đó là thông tin duy nhất mà chúng ta dựa vào để phân tích, xử lý và rút ra kết luận. Vì vậy, đòi hỏi thông tin đó phải đại diện trung thực và khách quan cho tập giá trị mà biến ngẫu nhiên X nhận. Để đạt được điều đó, trước tiên chúng ta phải tìm hiểu về lý thuyết mẫu.

II.2. LÝ THUYẾT MẪU

II.2.1. Một vài phương pháp lấy mẫu đơn giản

* Các quan sát độc lập hay các phép thử độc lập:

Các quan sát (phép thử) được tiến hành một cách độc lập với nhau, kết quả của quan sát (phép thử) này không phụ thuộc vào kết quả của quan sát (phép thử) khác và cũng không ảnh hưởng đến khả năng xảy ra kết quả của quan sát (phép thử) khác.

* Các phép thử lặp là các phép thử được tiến hành trong các điều kiện hoàn toàn như nhau.

* Lấy mẫu ngẫu nhiên có hoàn lại: Rút ngẫu nhiên từ một tập nào đó ra một phần tử, ghi lại các số đặc trưng cần thiết từ phần tử đó, sau đó trả nó trở lại tập ban đầu trước khi rút tiếp ngẫu nhiên lần sau.

* Lấy mẫu ngẫu nhiên không hoàn lại: Tương tự như trên, chỉ khác ở chỗ các phần tử được rút ra sẽ không trả lại tập ban đầu.

Ví dụ II.1:

5 xạ thủ, mỗi người bắn một viên vào bia. Đó là 5 phép thử hoặc 5 quan sát độc lập.

Một xạ thủ bắn 5 viên vào bia. Đó là 5 phép thử lặp và cũng là 5 phép thử độc lập.

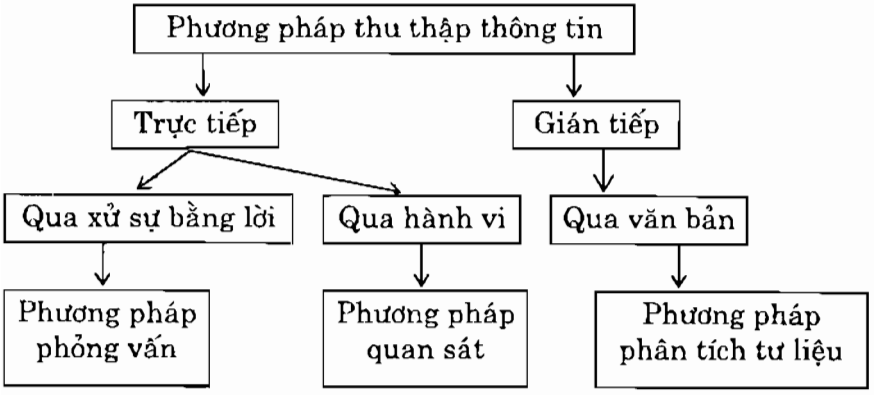
Quay giải nhất của xổ số gồm 5 chữ số bằng cách dùng một bộ số, một bàn quay và quay 5 lần. Đó là lấy mẫu có hoàn lại 5 lần.

Kiểm tra một số hộp sữa để đánh giá chất lượng của lô sữa. Những hộp được lấy ngẫu nhiên ra để kiểm tra chính là lấy mẫu không hoàn lại.

Lấy mẫu là một vấn đề rất quan trọng và phong phú trong nghiên cứu thống kê. Ngoài các phương pháp lấy mẫu chung dùng trong khoa học thống kê, đối với mỗi lĩnh vực riêng biệt như nông nghiệp, lâm nghiệp, sinh học, địa chất, xã hội học... lại có những phương pháp lấy mẫu mang tính đặc thù riêng.

Giáo trình này giới thiệu phương pháp lấy mẫu trong nghiên cứu xã hội học.

Có thể khái quát các phương pháp thu thập thông tin trong xã hội bằng sơ đồ sau:



Phỏng vấn là phương pháp thu thập thông tin chủ yếu. Thông qua các câu hỏi có mục đích và trả lời của các cá nhân chọn lọc, người ta thu thập thông tin về các hiện tượng xã hội. Bằng các hình thức phỏng vấn khác nhau, trong một thời gian tương đối ngắn, người ta có thể thu một lượng thông tin lớn về đối tượng nghiên cứu. Hình thức phỏng vấn chính là phỏng vấn miệng: điều tra viên tiếp xúc trực tiếp với người được phỏng vấn và thông tin được thu thập bằng lời. Nếu việc trả lời được thực hiện bằng cách điền vào một bảng câu hỏi in

sẵn thì được gọi là phỏng vấn viết. Phỏng vấn có thể có hiệu quả cao hơn nếu được kết hợp với các phương pháp khác như quan sát, thí nghiệm hoặc thảo luận.

Điều tra bằng phỏng vấn có hai phần chính: cách lập bảng câu hỏi và kỹ thuật viên điều tra. Cả hai phần đều đóng vai trò quan trọng trong việc nâng cao chất lượng thông tin thu thập được. Bảng câu hỏi cần được xây dựng sao cho đầy đủ, hợp lý, logic và phù hợp tâm lý. Ở một số nước, người ta xây dựng những thư viện mẫu về các câu hỏi điển hình cho các đối tượng khác nhau và cho những nội dung khác nhau. Trong khi đó, kỹ thuật viên điều tra đòi hỏi có nghiệp vụ thống kê và xã hội học, có kỹ thuật tâm lý thực hành, có năng khiếu về giao tiếp với nhiều loại đối tượng.

Nhưng đối tượng điều tra trong xã hội học quá lớn, chúng ta không có đủ điều kiện về thời gian, tài chính để điều tra mọi đối tượng, vì vậy vấn đề đặt ra là chọn những đối tượng nào để điều tra và cách chọn ra sao. Đó chính là vấn đề chọn mẫu trong điều tra xã hội học. Có 3 cách chọn mẫu như sau:

– Chọn mẫu với xác suất đều.

– Chọn mẫu với xác suất không đều.

– Chọn mẫu theo nhóm trội.

** Chọn mẫu với xác suất đều*

Ứng với mỗi đơn vị điều tra có một hay nhiều tiêu thức điều tra. Người ta thường phân loại các tiêu thức điều tra và giải quyết các vấn đề chọn mẫu theo hai cách sau:

– Tiêu thức về kết cấu xã hội của các cá nhân thuộc tổng thể điều tra như tuổi, giới tính, nghề nghiệp, địa vị, trình độ văn hóa... được chọn làm tiêu thức chọn mẫu. Người ta chọn

mẫu sao cho cơ cấu theo các tiêu thức này tính trên mẫu chọn trùng với cơ cấu của chúng đã biết trên tổng thể.

– Tiêu thức về các đặc tính riêng cần điều tra được chọn làm tiêu thức chọn mẫu.

Tính đại diện của mẫu thể hiện ở chỗ tiêu thức điều tra đo lường trên mẫu có thể suy rộng cho tổng thể, cũng như phân phối của nó trên mẫu có thể suy ra phân phối của tiêu thức đó trên tổng thể với độ tin cậy nào đó. Mẫu đại diện là hình ảnh thu nhỏ của tổng thể chung một cách tương đối trung thực.

** Chọn mẫu với xác suất không đều*

Trên các lĩnh vực xã hội, sự không đồng đều về quy mô của hiện tượng là phổ biến, do đó việc thu thập thông tin dựa vào các mẫu đồng loạt sẽ hạn chế tính đại diện, chẳng hạn các hiện tượng xã hội, văn hóa, tinh thần không xuất hiện đồng đều theo thời gian mà thường rộ lên ở một số thời điểm nhất định. Chọn mẫu theo thời gian với xác suất không đều tùy theo quy mô xuất hiện các hiện tượng cho các mẫu đại diện tốt hơn các mẫu chọn theo xác suất đều. Ví dụ: hiện tượng mê tín dị đoan tập trung vào dịp lễ hội, vào nơi có lễ hội. Những người nhiễm HIV tập trung ở những gái mại dâm, người tiêm chích ma tuý...

Chẳng hạn việc chọn hộ gia đình ở thành phố có thể thực hiện theo hai cách:

– Chọn mẫu theo hệ thống quản lý hành chính gồm các cấp: quận – phường – tổ dân phố – hộ gia đình. Nếu việc phân chia các đơn vị hành chính tương đối đều về số dân thì mẫu nhiều cấp với xác suất đều là hợp lý.

– Chọn mẫu theo kết cấu tự nhiên của dân cư thành phố,

gồm 3 cấp: các đường phố – các số nhà – các hộ gia đình (khu tập thể coi là một phố). Do quy mô của đơn vị mẫu ở mỗi cấp không đều nhau: số lượng số nhà của một đường phố, số lượng hộ gia đình của mỗi số nhà khác nhau, cho nên hợp lý nhất là chọn mẫu với xác suất không đều tỷ lệ với các số lượng đó; chỉ có chọn đường phố là theo xác suất đều.

* *Điều tra nhóm trội*

Điều tra nhóm trội là một dạng điều tra trọng điểm, nhưng với những điều kiện nhất định có khả năng suy rộng cho toàn bộ tổng thể. Chẳng hạn, nghiên cứu quỹ thời gian rỗi của nhóm người có điều kiện sống cao, có giao lưu văn hóa mạnh, có trình độ văn hóa nhất định... ta có thể suy rộng (có điều chỉnh) ra cho toàn bộ dân cư với khoảng thời gian đủ dài để nâng cao mức sống vật chất và tinh thần của toàn thể lên mức hiện tại của nhóm trội.

II.2.2. Mẫu ngẫu nhiên

Tiến hành n quan sát độc lập về biến ngẫu nhiên X nào đó. Ta gọi X_i là việc quan sát lần thứ i về biến ngẫu nhiên X. Khi đó $(X_1, X_2,..., X_n)$ được gọi là mẫu ngẫu nhiên, n được gọi là cỡ mẫu. Như vậy, mẫu ngẫu nhiên cỡ n thực chất là n biến ngẫu nhiên độc lập, cùng phân phối như biến ngẫu nhiên X.

Ta gọi x_i là kết quả của quan sát ở lần thứ i. Khi đó $(x_1, x_2, ..., x_n)$ là n giá trị cụ thể mà X nhận và ta quan sát được, hay là một giá trị cụ thể mà mẫu ngẫu nhiên $(X_1, X_2,..., X_n)$ nhận.

Từ nay về sau khi nói rằng có một mẫu ngẫu nhiên cỡ n được rút ra từ biến ngẫu nhiên X, ta sẽ hiểu đó là n biến ngẫu nhiên độc lập, cùng phân phối nếu ta không quan tâm đến kết quả cụ thể đã quan sát được mà ta muốn nghiên cứu

tính chất chung của mẫu, của các đặc trưng mẫu; còn ta sẽ hiểu đó là n giá trị cụ thể quan sát được, nếu ta quan tâm đến kết quả cụ thể và cần những tính toán cụ thể.

Như vậy, chúng ta có trong tay một mẫu ngẫu nhiên. Đó là thông tin duy nhất để dựa vào đó và dùng các phương pháp, kết quả của Thống kê Toán học để phân tích và rút ra các kết luận cần thiết.

Do vậy, đòi hỏi mẫu phải đại diện trung thực và khách quan cho hiện tượng hoặc cho tập giá trị của đại lượng mà ta đang nghiên cứu.

Giả sử $(x_1, x_2, ..., x_n)$ là n giá trị quan sát được. Bạn đọc cần phân biệt khi nào dãy số liệu thu được là mẫu đại diện, khi nào không phải là mẫu đại diện, khi nào là dãy số liệu đầy đủ và chính xác? Cách xử lý và rút ra kết luận trong các trường hợp đó ra sao? Tại sao phải dùng mẫu đại diện? v.v...

Để giúp bạn đọc phân biệt, tác giả nêu ra một ví dụ cụ thể sau:

Trong cuộc bầu cử Tổng thống ở một nước, mỗi cử tri tự tay bỏ phiếu bầu của mình. Kết thúc giờ bỏ phiếu Ban bầu cử toàn quốc có được tập các phiếu bầu (tập này gồm hàng chục triệu phiếu tương ứng với số cử tri của nước đó).

Sau khi kiểm phiếu người ta thu được tỷ lệ phiếu bầu mà mỗi ứng cử viên thu được. Toàn bộ phiếu bầu trên không phải là mẫu đại diện mà là tập thông tin đầy đủ và chính xác. Tỷ lệ phiếu bầu của từng ứng cử viên là con số chính xác. Mặc dù để có được các thông tin đầy đủ và chính xác như trên sẽ rất tốn kém, nhưng dù tốn kém vẫn phải làm và không có cách nào khác thay thế được.

Song trước khi bầu cử, các ứng cử viên có một vài tháng để vận động tranh cử. Bản thân các ứng cử viên và các tổ

chức, các đảng phái cũng muốn biết được thái độ cử tri đối với từng ứng cử viên ra sao, sau mỗi lần vận động, thái độ cử tri thế nào v.v... Để có được thông tin, trong trường hợp này người ta không thể tổ chức bầu cử thử. Vì vậy, các phóng viên, các điều tra viên đã phỏng vấn một số cử tri được chọn ngẫu nhiên (theo quy tắc chọn mẫu) và họ có được một lượng thông tin. Lượng thông tin này là mẫu đại diện. Dựa trên mẫu này, bằng các phương pháp và kết quả của thống kê ứng dụng người ta đánh giá được tỷ lệ phiếu bầu đối với các ứng cử viên. Chẳng hạn, ứng cử viên A thu được 36%, ứng cử viên B thu được 30%, ứng cử viên C thu được 18%, v.v... Các con số này chỉ cho ta thông tin là: ứng cử viên A đang dẫn đầu, ứng cử viên C đứng thứ 3, chênh lệch giữa ứng cử viên A với ứng cử viên tiếp theo là 6% (6 điểm), chưa có ứng cử viên nào vượt được quá bán...

Các ứng cử viên, các tổ chức, đảng phái cũng chỉ cần các thông tin quan trọng trên. Chắc chắn không có ai lấy tỷ lệ ước lượng trên làm tỷ lệ phiếu bầu thực mà họ sẽ thu được.

Độ tin cậy của các thông tin rút ra trên sẽ phụ thuộc rất nhiều vào mẫu đại diện được chọn như thế nào, có phản ánh trung thực ý kiến của cử tri toàn nước đó hay không. Thực tế ở một số nước tổ chức bầu Tổng thống hoặc Thủ tướng trong những năm gần đây như ở Mỹ, Anh, Pháp, Đức... các thông tin thu được trước khi bầu cử thực thường phản ánh đúng trong kết quả bầu cử.

Bạn đọc hãy chỉ ra các tình huống trong thực tế mà người ta xử lý dựa trên mẫu đại diện (càng nhiều, càng cụ thể càng tốt).

II.2.3. Cách thu gọn và biểu diễn số liệu

Giả sử $(x_1, x_2,..., x_n)$ là n giá trị thu được. Ta gọi dãy số liệu mẫu này là mẫu nguyên thủy hay mẫu ban đầu. Từ mẫu nguyên thủy người ta tìm cách thu gọn để tiện lưu giữ và sử dụng, tìm cách biểu diễn dưới dạng biểu đồ để cho ta hình ảnh về toàn bộ tập số liệu mẫu, từ đó dễ rút ra các xu thế, các tính chất của mẫu, dễ so sánh,...

a) Mẫu thu gọn

Vì trong n giá trị thu được có thể có nhiều giá trị trùng nhau nên ta gộp các giá trị trùng nhau lại và kể đến số lần mẫu nhận giá trị đó, khi đó ta có:

$x_{(1)}$	$x_{(2)}$	$x_{(k)}$
m_1	m_2	m_k

$$\sum_{i=1}^{k} m_i = n$$

trong đó: m_i là số lần n giá trị mẫu nhận $x_{(i)}$; k là số các giá trị mẫu khác nhau, ta gọi là cỡ mẫu thu gọn. Nếu $k = n$ thì $m_i = 1\ \forall\ i$.

Để đơn giản cách viết, thay cho viết $x_{(i)}$ ta viết x_i, nhưng cần lưu ý rằng giá trị x_i trong mẫu thu gọn nói chung không phải là x_i trong mẫu ban đầu.

b) Mẫu thu gọn dạng khoảng

Nếu cỡ mẫu lớn ta có thể thu gọn số liệu dưới dạng khoảng. Chia khoảng giá trị mẫu thành các khoảng con $[a_i, a_{i+1})$ với các điểm chia a_i. Gọi m_i là số giá trị mẫu rơi vào khoảng con $[a_i, a_{i+1})$. Rõ ràng $\Sigma m_i = n$.

Việc chọn các điểm chia a_i, độ dài các khoảng con, kể đầu mút nào vào khoảng con v.v... là tùy thuộc vào người xử lý số

liệu. Thông thường ta chọn độ dài các khoảng con bằng nhau và kín ở mút trái.

Khi tính toán, nếu số liệu cho dưới dạng khoảng ta phải thay khoảng đó bởi một điểm đại diện, nghĩa là ta đã làm một sự xấp xỉ, do đó nếu độ dài của khoảng con càng lớn thì sai số trong xấp xỉ càng lớn. Như vậy, mẫu ở dạng thu gọn $(x_{(i)}, m_i)$ sẽ cho ta các kết quả tính toán chính xác, còn mẫu ở dạng khoảng cho kết quả tính toán đòi hỏi sự hiệu chỉnh.

c) Biểu đồ tần suất, đa giác tần suất

Từ mẫu thu gọn, tại mỗi điểm $x_{(i)}$ ta chấm một điểm với tung độ là $\frac{m_i}{n}$ (tần suất $\frac{m_i}{n}$ thường được gọi là tỷ lệ phần trăm %). Tập điểm nhận được cho ta biểu đồ tần suất. Nối các điểm liên tiếp với nhau ta nhận được đa giác tần suất.

Nhìn vào đa giác tần suất ta được hình ảnh về tập số liệu mẫu, những nét đặc trưng của chúng...

Nếu số liệu được cho dưới dạng khoảng, ta chọn mỗi khoảng một điểm đại diện, thông thường là điểm giữa của mỗi khoảng và số liệu lại được đưa về dạng thu gọn $(x_{(i)}, m_i)$ như trên.

Nếu có nhiều mẫu khác nhau, để dễ so sánh ta biểu diễn trên cùng một biểu đồ nhưng với các ký hiệu điểm khác nhau (hoặc màu sắc khác nhau).

d) Biểu đồ hình chữ nhật

Trên mỗi đặc trưng, mỗi tiêu thức, mỗi khoảng ta vẽ một hình chữ nhật với chiều cao là $\frac{m_i}{n}$ (hoặc tỷ lệ phần trăm).

Tập các hình chữ nhật nhận được cho ta biểu đồ chữ nhật. Biểu đồ chữ nhật cho ta hình ảnh về tập số liệu mẫu và

sự so sánh giữa các tiêu thức (giữa các khoảng). Nếu có vài mẫu số liệu khác nhau ta có thể biểu diễn các hình chữ nhật tương ứng cùng một tiêu thức cạnh nhau nhưng với màu sắc khác nhau và các tiêu thức khác nhau của cùng một mẫu thì cùng một màu. Khi đó, nhìn các hình chữ nhật cùng màu cho ta sự so sánh giữa các tiêu thức, nhìn vào cùng một tiêu thức với các hình chữ nhật màu sắc khác nhau sẽ cho ta sự so sánh giữa các mẫu khác nhau của tiêu thức này.

e) Biểu đồ hình quạt

Thay cho việc biểu diễn bằng hình chữ nhật, ta biểu diễn tỷ lệ phần trăm các đặc trưng (tiêu thức) bằng các hình quạt trong một vòng tròn với tỷ lệ diện tích hình quạt so với diện tích hình tròn chính là tỷ lệ phần trăm của tiêu thức đó.

Ví dụ II.2: Theo báo cáo đánh giá kết quả thi và công tác tổ chức thi học kỳ tại trường Đại học Đại cương (Báo cáo của Ban Giám hiệu Trường Đại học Đại cương thuộc Đại học Quốc gia Hà Nội (ngày 24/4/1998)), ta có các số liệu sau (số liệu ở đây là số liệu điều tra đầy đủ, chính xác, không phải số liệu mẫu. Do đó, các kết luận rút ra là chân thực đối với khóa II Đại học Đại cương. Qua tập số liệu điều tra đầy đủ, chính xác này chúng ta làm quen với việc thu gọn, biểu diễn và phân tích tập số liệu quan sát thu được).

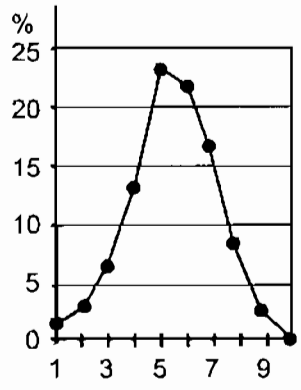
1) Kết quả thi học kỳ của sinh viên khóa II Đại học Đại cương (ĐHĐC):

Tổng số bài thi 156157 bài, (26 học phần/ 3 học kỳ), tổng số sinh viên: 6823. Như vậy, ta có 156157 điểm thi. Số liệu được thu gọn trong 2 cột đầu, cột 3 là tỷ lệ phần trăm hoặc tần suất $\frac{m_i}{n}$.

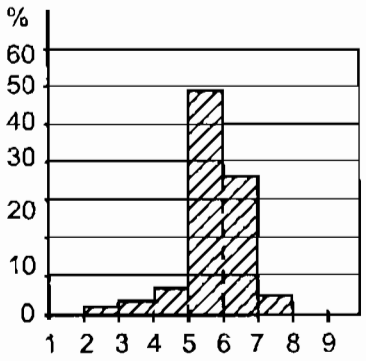
Điểm thi	Số bài thi (m_i)	Phần trăm = tần suất $\frac{m_i}{n}$
0	2828	0,0181 = 1,81%
1	2436	0,0156 = 1,56%
2	4328	0,0277 = 2,77%
3	10059	0,0644 = 6,44%
4	200677	0,1285 = 12,85%
5	36792	0,2356 = 23,56%
6	34466	0,2207 = 22,07%
7	26079	0,1670 = 16,70%
8	12885	0,0825 = 8,25%
9	5015	0,0321 = 3,21%
10	1202	0,0078 = 0,78%
Σ	156157	1,0000 = 100%

156157 số liệu trên có thể được thu gọn dưới dạng khoảng như sau:

Điểm thi	Số bài thi (m_i)	Phần trăm = tần suất $\frac{m_i}{n}$
[0, 3] (Phải thi lại)	19651	0,1258 = 12,58%
[4, 6]	91325	0,5848 = 58,48%
[7, 10] (Loại khá, giỏi)	45181	0,2893 = 28,94%



a)



b)

Hình II.1.

a) Biểu đồ tần suất điểm thi khóa II ĐHĐC

b) Biểu đồ điểm trung bình 3 học kỳ khóa II ĐHĐC

2) Điểm trung bình của sinh viên khóa II – ĐHĐC (sau 3 học kỳ):

Tổng kết sau 3 học kỳ mỗi sinh viên sẽ có một điểm trung bình, ta dùng điểm đó để xét chuyển thẳng và xét cấp chứng chỉ ĐHĐC (Điểm trung bình = $\left(\sum_{i=1}^{26} w_i x_i\right) / \sum_{i=1}^{26} w_i$, trong đó x_i là điểm thi của môn thứ i với w_i đơn vị học trình. 3 học kỳ có 26 môn thi (26 học phần). Trong số 6629 sinh viên có điểm thi trung bình cả 3 học kỳ, thu gọn số liệu dạng khoảng ta có:

Điểm trung bình	Số sinh viên (m_i)	Tần suất $\frac{m_i}{n}$ = phần trăm
[2; 3)	44	0,0064 = 0,64%
[3; 4)	55	0,0081 = 0,81%
[4; 5)	353	0,0517 = 5,17%
[5; 6)	3623	0,5310 = 53,1%
[6; 7)	2256	0,3310 = 33,1%
[7; 8)	288	0,0422 = 4,22%
[8; 9)	10	0,0015 = 0,15%
[9; 10]	0	0 = 0
Tổng	6629	1,000 = 100%

Nhìn vào các bảng số liệu thu gọn và biểu đồ nhận được ở trên ta thấy:

– Phân bố điểm thi là phân bố gần chuẩn hơi lệch về phía điểm khá, giỏi.

– Tỷ lệ sinh viên có điểm trung bình từ 5 đến dưới 6 là cao nhất (53%).

– Tỷ lệ sinh viên đạt tiêu chuẩn điểm chuyển thẳng (≥ 6) là 37,5% (2554 sinh viên).

– Tỷ lệ sinh viên tốt nghiệp ĐHĐC loại khá giỏi (≥ 7) là 4,37%.

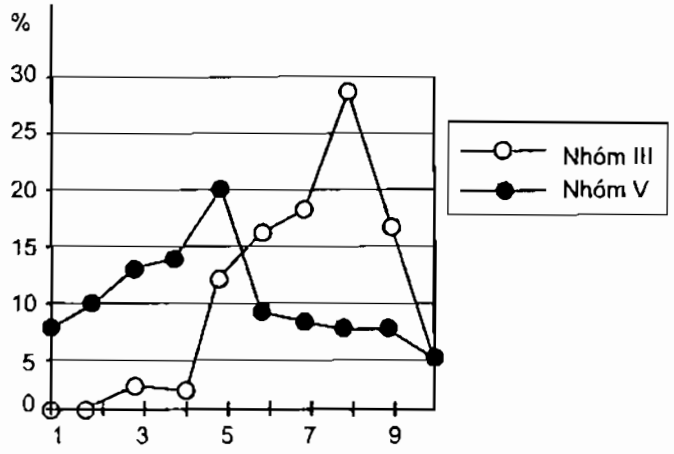
3) Điểm thi của một vài môn cụ thể được thể hiện qua biểu đồ tần suất:

– Môn Toán D (nhóm ngành III) Đại học Khoa học Tự nhiên (TN).

– Môn Toán E (nhóm ngành V) Đại học Khoa học Xã hội và Nhân văn (XHNV).

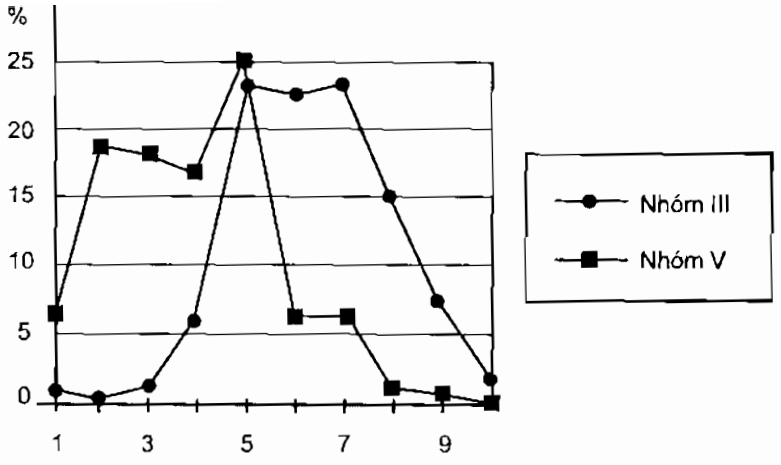
– Môn Toán D (nhóm ngành III) và Toán E (nhóm ngành V) Đại học Sư phạm (SP).

– Môn Thống kê xã hội học ở các nhóm ngành VI SP, XHNV và nhóm ngành VII Ngoại ngữ (NN).



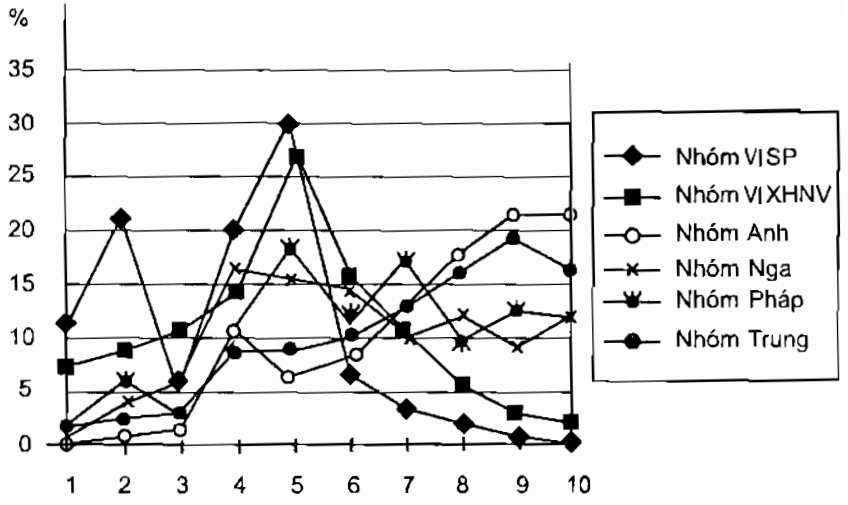
Hình II.2.

Điểm thi môn Toán D (nhóm ngành III) TN và E (nhóm ngành V) XHNV



Hình II.3.

Điểm thi môn Toán D (nhóm ngành III) TN và E (nhóm ngành V) SP



Hình II.4. Điểm thi môn Thống kê xã hội học

Nhìn vào các biểu đồ trên ta thấy:

– Sinh viên nhóm ngành V học môn Toán tồi hơn sinh viên nhóm ngành III (ở cả 2 trường Tổng hợp cũ và Trường Sư phạm).

– Ở nhóm ngành III tỷ lệ sinh viên có điểm ≤ 4 thấp, tỷ lệ điểm ≥ 6 khá cao.

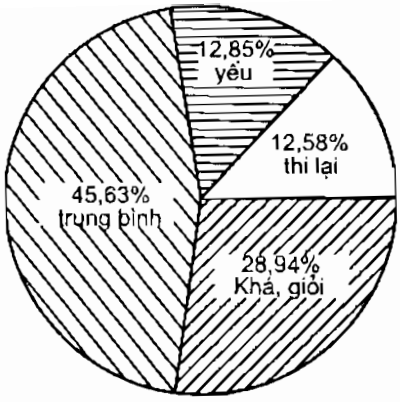
– Ở nhóm ngành V tỷ lệ điểm ≤ 4 lại khá cao, còn tỷ lệ điểm ≥ 6 lại rất thấp.

– Kiến thức toán, tư duy toán học của sinh viên nhóm ngành V (khóa II) rõ ràng có lỗ hổng lớn (thực tế có không ít sinh viên nhóm này làm toán cấp II không thạo, không ít sinh viên không học môn Toán khi học ở cấp III (THPT) (mặc dù họ vẫn có điểm tổng kết môn Toán từng năm và điểm thi tốt nghiệp đạt yêu cầu)).

– Nhìn vào biểu đồ của môn Thống kê xã hội học ta thấy chất lượng đào tạo phụ thuộc rất nhiều vào chất lượng đầu vào. Nhóm ngành VI (SP và XHNV) phân bố điểm lệch về phía điểm thấp, thậm chí quá thấp. Trong khi đó, ở nhóm các khoa Ngoại ngữ lại có điểm lệch rất nhiều sang phía điểm cao (xấp xỉ 40% – 50% điểm giỏi (≥ 8)). (Môn này đều cùng do một nhóm các thầy cô thuộc khoa Toán – Cơ – Tin học Đại học Khoa học Tự nhiên dạy).

4) Điểm thi học kỳ của sinh viên khóa II – ĐHĐC được chia làm 4 loại: Phải thi lại (<4 điểm), loại yếu (≤4 điểm), loại trung bình (5–6 điểm) và loại khá giỏi (≥7 điểm) (xem ở bảng 1 trong ví dụ này).

Biểu diễn kết quả trên dưới dạng biểu đồ hình tròn (hình II.5).



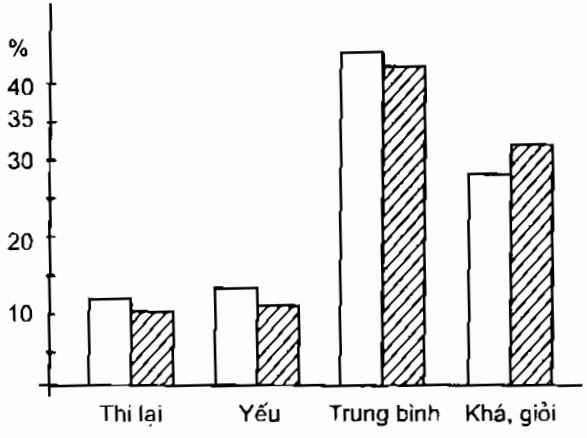
Hình II.5. Điểm thi học kỳ khóa II ĐHĐC

Thống kê kết quả thi học kỳ I của sinh viên khóa III – ĐHĐC (35565 bài thi) ta có sự phân loại sau:

Phải thi lại: 10,3% Yếu: 11,77%

Trung bình: 43,69% Khá giỏi: 34,24%

Hãy biểu diễn trên biểu đồ chữ nhật để tiện so sánh kết quả thi học kỳ của khóa II và kỳ I khóa III – ĐHĐC.



Hình II.6.

Miền không gạch – tương ứng với khóa II; Miền gạch – tương ứng với kỳ I khóa III

Nhìn vào biểu đồ chữ nhật ta thấy kết quả thi của kỳ I khóa III so với khóa II đã có sự tiến bộ. Điểm thi lại, điểm yếu, điểm trung bình đều giảm, còn điểm khá giỏi lại tăng lên.

f) Biểu diễn dãy số liệu hai chiều

Giả sử ta quan sát đồng thời hai biến X, Y nào đó, mỗi quan sát cho ta một cặp giá trị (x, y). Giả sử n quan sát độc lập cho ta n giá trị (x_i, y_i), i = 1,2,...n.

Nếu chấm n cặp giá trị này lên mặt phẳng tọa độ vuông góc ta sẽ nhận được một tập các điểm. Tập điểm này cho ta xu thế biến thiên và dạng phụ thuộc giữa 2 biến (xem phần IV cuối chương này).

Trong n cặp giá trị này có thể có cặp trùng nhau, khi đó có thể thu gọn dãy số liệu dưới dạng mẫu thu gọn như trường hợp a). Ta có:

$$\begin{cases} (x_i, y_i) \\ m_i\ , \quad \sum_{i=1}^{k} m_i = n \end{cases}$$

k là số cặp giá trị khác nhau. Nếu $m_i = 1$, $\forall$ i thì k = n.

Mẫu thu gọn có thể biểu diễn bởi 3 hàng hoặc 3 cột nhưng lưu ý là các giá trị của x và y phải tương ứng với nhau:

x_i:	x_1	x_2	x_k
y_i:	y_1	y_2	y_k
m_i:	m_1	m_2	m_k

Ngay trong mẫu thu gọn các giá trị $x_1, x_2, ..., x_k$ vẫn có thể có những giá trị trùng nhau, chẳng hạn có r giá trị khác nhau $x_{(1)}, x_{(2)}, ..., x_{(r)}$. Tương tự, trong k giá trị $y_1, y_2, ..., y_k$ có s giá trị khác nhau: $y_{(1)}, y_{(2)}, ..., y_{(s)}$. Khi đó, ta có thể biểu diễn mẫu thu gọn dưới dạng bảng 2 lối vào gồm r hàng và s

cột, do đó gồm r.s ô chữ nhật, trong mỗi ô này (chẳng hạn ô (i,j) – ô ở hàng i và cột j) ta ghi số giá trị trong n giá trị mà biến X nhận $x_{(i)}$, biến Y nhận giá trị $y_{(j)}$.

Con số này ta ký hiệu là n_{ij}. Thực chất n_{ij} sẽ là giá trị m_i nào đó hoặc là 0.

Biểu diễn dãy số liệu hai chiều dưới dạng mẫu thu gọn (3 cột hoặc 3 hàng) và dưới dạng bảng 2 lối vào là tương đương nhau.

Chẳng hạn, ta có n = 10 số liệu quan sát từ 2 biến X và Y như sau:

(1;5) (2;4) (2;5) (3;4) (4;5) (3;7) (1;5) (3;4) (4;5) (4;5)

Mẫu thu gọn sẽ là:

$$\begin{cases} (1;5)\ (2;4)\ (2;5)\ (3;4)\ (4;5)\ (3;7) \\ 2 \quad 1 \quad 1 \quad 2 \quad 3 \quad 1 \end{cases} \Rightarrow k = 6$$

Ta có các cách biểu diễn sau:

x_i	1	2	2	3	3	4
y_i	5	4	5	4	7	5
m_i	2	1	1	2	1	3

x_i	y_i	m_i
1	5	2
2	4	1
2	5	1
3	4	2
3	7	1
4	5	3
	Σ	10

X \ Y	4	5	7
1	0	2	0
2	1	1	0
3	2	0	1
4	0	3	0

Ở đây: n = 10; k = 6; r = 4; s = 3

II.2.4. Các đặc trưng mẫu

Giả sử ta nghiên cứu biến ngẫu nhiên X.

Ký hiệu $\mu = EX$, $\sigma^2 = DX$.

Các số đặc trưng của X gọi là các số đặc trưng lý thuyết. (Các số này ta chưa biết và đang cần tìm).

Giả sử $(X_1, X_2,..., X_n)$ là mẫu ngẫu nhiên rút ra từ X.

a) Kỳ vọng mẫu (Giá trị trung bình mẫu)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Do $X_1, X_2,..., X_n$ là các biến ngẫu nhiên độc lập, cùng phân phối như X nên $\overline{X}$ là một biến ngẫu nhiên. Do đó, theo tính chất của kỳ vọng và phương sai ta có:

$$E\overline{X} = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{n.EX}{n} = \mu.$$

$$D\overline{X} = \frac{1}{n^2}\sum_{i=1}^{n} DX_i = \frac{n.DX}{n^2} = \frac{DX}{n} = \frac{\sigma^2}{n}.$$

b) Phương sai mẫu

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2 .$$

Ta có:

$$Es^2 = \frac{1}{n}.E\sum_{i=1}^{n}\left\{(X_i - \mu) - (\bar{X} - \mu)\right\}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}E(X_i - \mu)^2 - E(\bar{X} - \mu)^2$$

$$= \frac{1}{n}.n.DX - \frac{DX}{n} = \frac{n-1}{n}.\sigma^2$$

Đôi khi thay cho việc tính s^2 ta tính $\hat{s}^2$ trong đó:

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n}{n-1}.s^2 .$$

Khi đó

$$E(\hat{s}^2) = \frac{n}{n-1}.Es^2 = \frac{n}{n-1}.\frac{n-1}{n}.\sigma^2 = \sigma^2 .$$

Như vậy, về mặt trung bình thì $\bar{X}$ và $\hat{s}^2$ đúng bằng kỳ vọng và phương sai lý thuyết mà ta đang cần tìm.

c) Phân bố của $\bar{X}$ và s^2

– Nếu $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên được rút ra từ biến ngẫu nhiên chuẩn $N(\mu, \sigma^2)$ thì:

+ $n\bar{X}$ cũng có phân phối chuẩn $N(n\mu, n\sigma^2)$.

+ $\bar{X}$ có phân phối chuẩn $N(\mu, \frac{\sigma^2}{n})$.

+ $\frac{ns^2}{\sigma^2}$ có phân phối χ^2_{n-1}.

+ $\bar{X}$ và s^2 độc lập với nhau (và ngược lại):

$$t = \frac{\bar{X} - \mu}{s}\sqrt{n-1} \text{ hoặc } t = \frac{\bar{X} - \mu}{\hat{s}}\sqrt{n}$$

có phân phối Student với n – 1 bậc tự do.

– Nếu $(X_1, X_2, ..., X_n)$ là mẫu từ biến ngẫu nhiên chuẩn $N(\mu_1, \sigma_1^2)$, còn $(Y_1, Y_2, ..., Y_m)$ là mẫu từ biến ngẫu nhiên chuẩn $N(\mu_2, \sigma_2^2)$, độc lập với mẫu trên thì:

$$\bar{X} \pm \bar{Y} = N(\mu_1 \pm \mu_2; \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m})$$

$$t = \frac{\left[(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)\right]\sqrt{(n+m-2).n.m}}{\sqrt{(n+m).(n.s_x^2 + m.s_y^2)}}$$

có phân phối Student với (n + m – 2) bậc tự do, trong đó s_x^2, s_y^2 là hai phương sai mẫu tương ứng với mẫu X và mẫu Y.

II.2.5. Cách tính $\bar{X}$ và s^2

Cho mẫu ngẫu nhiên $(x_1, x_2, ..., x_n)$ hoặc mẫu thu gọn

$$\begin{cases} x_1 & x_2 & \dots & x_k \\ m_1 & m_2 & \dots & m_k \end{cases}; \qquad \sum_{i=1}^{k} m_i = n$$

Khi đó:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} m_i x_i ; \qquad s^2 = \frac{1}{n}\sum_{i=1}^{k} m_i x_i^2 - \bar{X}^2$$

Nếu các giá trị mẫu x_i không gọn (khó tính nhẩm) nhưng chúng lại cách đều nhau thì ta có thể rút gọn số liệu bằng phép biến đổi tuyến tính:

$$u_i = \frac{x_i - x_0}{h}$$

(Có nghĩa là ta thay đổi gốc tính và đơn vị tính). Chọn hằng số x_0 và h căn cứ vào dãy số liệu cụ thể, thông thường ta lấy giá trị x_0 là giá trị x_i ở quãng giữa, còn h là độ dài khoảng cách đều giữa các giá trị x_i. Ta lập bảng tính như sau:

Bảng tính $\overline{X}$, s^2 đối với mẫu thu gọn:

x_i	m_i	m_ix_i	$m_ix_i^2$
⋮	⋮	⋮	⋮
Σ	n	(*)	(*)

Bảng tính $\overline{u}$, s_u^2 đối với số liệu rút gọn u_i của mẫu thu gọn:

x_i	m_i	u_i	m_iu_i	$m_iu_i^2$
⋮	⋮	⋮	⋮	⋮
Σ	n		(*)	(*)

Mỗi bảng tính ta sẽ nhận được 2 tổng cần thiết ở vị trí (*).

Từ bảng tính trên ta tính được:

$$\overline{u} = \frac{1}{n}\sum_{i=1}^{k} m_iu_i\,,\quad s_u^2 = \frac{1}{n}\sum_{i=1}^{k} m_iu_i^2 - \overline{u}^2$$

Trở lại số liệu ban đầu ta có:

$$\overline{X} = x_0 + h.\overline{u}\,,\quad s_x^2 = h^2.s_u^2$$

Trong trường hợp tính toán thiếu sự trợ giúp của máy tính thì việc lập bảng tính và rút gọn số liệu vẫn rất cần thiết và có lợi.

Nếu mẫu được cho dưới dạng các khoảng, ta chọn mỗi khoảng một điểm đại diện, thông thường là điểm giữa khoảng, lúc đó ta lại nhận được mẫu dạng thu gọn.

Ta có bảng tính đối với mẫu cho theo khoảng như sau:

Các khoảng	m_i	Điểm đại diện x_i	u_i	m_iu_i	$m_iu_i^2$
⋮	⋮	⋮	⋮	⋮	⋮
Σ	n			(*)	(*)

Do sự phát triển chung của xã hội, máy tính bỏ túi đã trở thành công cụ tính toán thông dụng, thì việc rút gọn số liệu

để tính nhẩm có thể bỏ qua. Do nhu cầu ứng dụng thực tế người ta đã lập sẵn các chương trình để tính $\bar{X}$, s^2 ngay cả đối với máy tính đơn giản bỏ túi, chẳng hạn các máy Casio fx 500A, fx 500 MS, v.v... Dưới đây giới thiệu sơ lược cách sử dụng chương trình thống kê ở hai loại máy Casio bỏ túi này.

Đối với máy Casio fx 500A:

Mở chương trình: MODE [·]

Xóa sạch số liệu cũ: SHIFT SAC

Nhập số liệu dạng thu gọn:

x_1 [x] m_1 [DATA] x_2 [x] m_2 [DATA] ... x_k [x] m_k [DATA]

Kết quả: SHIFT [n] cho số quan sát n =

SHIFT [$\bar{x}$] cho số quan sát $\bar{X}$ =

SHIFT [σ_n] cho số quan sát s =

SHIFT [σ_{n-1}] cho số quan sát $\hat{s}$ =

Thoát khỏi chương trình: MODE [0]

Đối với máy Casio fx 500 MS:

Mở chương trình: MODE [2]

Xóa sạch số liệu cũ: SHIFT CLR [1] [=]

Nhập số liệu:

x_1 SHIFT [;] m_1 [DT] x_2 SHIFT [;] m_2 [DT] ... x_k SHIFT [;] m_k [DT]

Kết quả: SHIFT [S–Sum] [3] [=] cho n =

SHIFT [S–Var] [1] [=] cho $\bar{X}$ =

SHIFT [S–Var] [2] [=] cho s =

SHIFT [S–Var] [3] [=] cho $\bar{s}$ =

Thoát khỏi chương trình: MODE [1]

Nhận xét: Qua việc trình bày trên, nếu xuất phát từ mẫu ban đầu, để tính $\bar{X}$ và s^2 ta nên thu gọn mẫu về dạng mẫu thu gọn sẽ tốt hơn việc thu gọn mẫu về dạng khoảng, vì ở dạng khoảng chúng ta đã tạo nên sai số trong tính toán khi thay mỗi khoảng bởi một điểm đại diện.

Ví dụ II.3: Các khu chế xuất, các công ty liên doanh đã thu hút khá nhiều lực lượng lao động là thanh niên, nhất là thanh niên nông thôn. Để đánh giá thu nhập của các công nhân này người ta đã điều tra ngẫu nhiên 50 công nhân. Kết quả cho thấy thu nhập theo tháng của từng người (đơn vị là nghìn đồng Việt Nam) như sau:

700 750 720 680 700 780 800 750 720 650 800
780 750 680 700 680 650 800 720 750 680 650
820 800 750 850 800 780 750 680 700 720 750
800 820 780 780 800 650 680 650 750 720 800
750 780 800 680 650 700

a) Hãy thu gọn số liệu mẫu trên thành mẫu thu gọn và mẫu dạng khoảng.

b) Tính $\bar{X}$, s, $\hat{s}$, s^2, $\hat{s}^2$ từ hai mẫu thu gọn. Nhận xét kết quả thu được.

Giải:

a) Mẫu thu gọn nhận được:

x_i	650	680	700	720	750	780	800	820	850
m_i	6	7	5	5	9	6	9	2	1

Ở đây: số giá trị khác nhau của mẫu hay cỡ mẫu thu gọn là k = 9; số quan sát hay cỡ mẫu n = 50.

Giá trị nhỏ nhất của mẫu là 650, giá trị lớn nhất của mẫu là 850.

Nếu gộp thành các khoảng với độ dài như nhau và bằng 50, ta có:

Khoảng thu nhập	[650 ; 700)	[700; 750)	[750; 800)	[800; 850]
Số công nhân	13	10	15	12

Nếu ta lấy độ dài của mỗi khoảng là 100 thì sẽ chỉ có 2 khoảng (vì độ dài khoảng càng rộng thì sai số tạo ra khi tính $\bar{X}$, s^2 càng lớn):

Khoảng thu nhập	[650 ; 750)	[750; 850]
Số công nhân	23	27

b) Dùng máy Casio fx 500 ta nhận được:

Đối với mẫu thu gọn:

$\bar{X}$ = 737,600 ; s = 55,011272 ; $\hat{s}$ = 55,569776 ;

$\hat{s}^2$ = 3088,000.

Đối với mẫu dạng khoảng (k = 4 khoảng) ta được:

$\bar{X}$ = 751 ; s = 55,892754 ; $\hat{s}$ = 56,460208 ;

$\hat{s}^2$ = 3187,755087

Đối với mẫu dạng khoảng (k = 2) ta được:

$\bar{X}$ = 754 ; s = 49,839743 ; $\hat{s}$ = 50,345743 ;

$\hat{s}^2$ = 2483,999982.

Ta thấy: khi tính $\bar{X}$, s^2 kết quả từ mẫu thu gọn là chính xác, còn kết quả từ hai mẫu dạng khoảng đã mắc sai số tính toán (giá trị $\bar{X}$ thì tăng lên, còn giá trị s, $\hat{s}$ thì giảm đi). Do đó, nếu không có yêu cầu gì đặc biệt, chúng ta nên thu gọn mẫu ở dạng thu gọn.

II.2.6. Sai số trong lấy mẫu

Khi lấy mẫu, do nhiều nguyên nhân khác nhau, sẽ không tránh khỏi những sai số trong các số liệu mẫu. Vì thế, trước khi dùng các phương pháp thống kê để phân tích, xử lý ta cần loại bỏ các sai số không đáng có ở trong mẫu đã cho.

Giả sử x là kết quả quan sát được, a là giá trị chân thực (giá trị đúng) của đại lượng ta quan sát, z là sai số trong lấy mẫu. Ta có:

Sai số trong lấy mẫu = Kết quả quan sát – Giá trị đúng

hay: $z = x - a$

Nói chung, vì a chưa biết nên sai số z cũng chưa biết.

Để thuận lợi cho việc xử lý ta phân loại các sai số như sau:

a) Sai số thô là sai số sinh ra do phạm vi các điều kiện cơ bản của việc lấy mẫu hoặc do sơ suất của người thực hiện, chẳng hạn, người kiểm tra cố ý chọn ra các sản phẩm tốt để kiểm tra khi đánh giá chất lượng của lô sản phẩm, hoặc kỹ thuật viên ghi nhầm kết quả thu được... Trong điều tra xã hội học, người điều tra vi phạm quy tắc chọn mẫu. Họ chọn mẫu một cách phiến diện, thiếu sự điều tra, tìm hiểu và phân tích trước khi lấy mẫu.

b) Sai số hệ thống là sai số do không điều chỉnh chính xác dụng cụ hoặc không thống nhất giữa các kỹ thuật viên về cách xác định một đại lượng nào đó... do vậy dẫn đến một loạt kết quả quan sát được bị lệch đi một lượng nào đó. Hoặc trong điều tra xã hội học, các điều tra viên không thống nhất câu hỏi phỏng vấn, không thống nhất các phương án trả lời... dẫn đến sự sai lệch thông tin giữa các điều tra viên.

c) Sai số ngẫu nhiên là sai số sinh ra do một số lớn các

nguyên nhân mà tác động của chúng nhỏ (ít) đến mức không thể tách riêng và tính riêng biệt cho từng nguyên nhân được.

Để hiểu rõ sai số ngẫu nhiên, xét hai ví dụ sau: Một xạ thủ bắn 100 viên đạn vào bia – Xạ thủ nhằm tâm để bắn. Với tài nghệ của mình thì phần lớn các viên đạn sẽ trúng vòng điểm 10, nhưng cũng sẽ có một số viên trúng vào vòng điểm 9, điểm 8,... Vẫn khẩu súng, loại đạn ấy, vẫn cùng tài nghệ ấy và xạ thủ cũng vẫn nhắm tâm để bóp cò, thế mà lại có viên trúng vòng điểm 9, lại có viên trúng vòng điểm 8. Hỏi lý do tại sao thì xạ thủ không giải thích được. Đó chính là sai số ngẫu nhiên.

Trong các cuộc thi thể thao, văn nghệ mà từng thành viên trong Ban Giám khảo sẽ đánh giá bằng cho điểm, chẳng hạn thi trượt băng nghệ thuật, thi hát,... hay trong việc đánh giá về một người nào đó bằng cách cho điểm. Trong những trường hợp như thế tất nhiên tiêu chuẩn và cách chấm điểm phải rõ ràng, thống nhất, loại trừ các động cơ cá nhân của người chấm điểm, nhưng không ai dám chắc mọi người đều chấm điểm như nhau cho một ứng viên nào đó. Có người cho hơi cao, lại có người cho hơi thấp hơn một chút. Đó là sai số ngẫu nhiên.

Trong 3 loại sai số trên, sai số thô, sai số hệ thống cần phát hiện sớm và khử bỏ ngay, còn sai số ngẫu nhiên không thể khử bỏ được trong mỗi lần lấy mẫu. Do đó từ nay về sau, khi các kết quả quan sát được đưa vào xử lý bằng phương pháp toán học ta sẽ giả thiết rằng chúng chỉ chứa các sai số ngẫu nhiên.

d) Phân phối của sai số ngẫu nhiên. Thông thường ta lấy luật phân phối chuẩn làm luật phân phối của sai số ngẫu nhiên. Điều đó thường khá phù hợp với thực nghiệm bởi lẽ

luật phân phối chuẩn phản ánh tính đối xứng của các sai số ngẫu nhiên: các sai số ngẫu nhiên có dấu hiệu khác nhau thường được gặp gần như nhau. Hơn nữa, luật phân phối chuẩn còn phản ánh tính tập trung của các sai số ngẫu nhiên: sai số ngẫu nhiên có trị số tuyệt đối bé thường gặp hơn các sai số có trị số tuyệt đối lớn.

Vậy $Z = N(0; \sigma^2)$

với σ^2 là độ chính xác của các lần lấy mẫu.

Ta có (xem chương I hoặc trang 171 Phụ lục I):

$$P\{a < Z < b\} = \Phi\left(\frac{b}{\sigma}\right) - \Phi\left(\frac{a}{\sigma}\right)$$

$$P\{|Z| < k\sigma\} = \Phi(k) - \Phi(-k)$$

trong đó: $\Phi(x)$ là phân phối chuẩn N(0;1).

Với: k = 2 ta có $P\{-2\sigma < Z < 2\sigma\} = 0,95$.

k = 3 ta có $P\{|Z| > 3\sigma\} \approx 0,0027$

Xác suất 0,0027 quá bé, cho nên ta xem như trong thực tế sai số ngẫu nhiên không thể vượt quá giới hạn $\pm 3\sigma$.

Trước khi chuyển sang nội dung chính của chương này, tác giả xin nhắc lại bài toán đầu tiên của chúng ta như sau: Ta muốn nghiên cứu một đặc trưng xã hội nào đó, ký hiệu là X, cụ thể là muốn tìm câu trả lời về hai đặc trưng quan trọng của X là giá trị trung bình và tỷ lệ, hay nói cách khác là kỳ vọng và xác suất, ta ký hiệu chúng là EX hay μ và $P(X \in A)$ hay p. Để trả lời, thông tin chúng ta có chỉ là mẫu đại diện, tức là thông tin ta có về biến ngẫu nhiên X vừa không đầy đủ lại còn chứa sai số. Dựa trên lượng thông tin như vậy không thể trông chờ một câu trả lời chính xác như là dựa trên lượng

thông tin đầy đủ, chính xác được. Tuy lượng thông tin còn thiếu và chứa sai số nhưng Thống kê Toán học cũng sẽ cho câu trả lời về giá trị trung bình thực μ và về tỷ lệ thực p của biến ngẫu nhiên X, và câu trả lời đó sẽ có độ tin cậy cao nhất theo nghĩa là có khả năng đúng cao nhất và khả năng bị sai thấp nhất.

II.3. MỘT VÀI ƯỚC LƯỢNG ĐƠN GIẢN

II.3.1. Ước lượng điểm cho kỳ vọng, Median, Mode, phương sai và xác suất

Giả sử θ là tham ẩn cần ước lượng; $X_1,..., X_n$ là mẫu ngẫu nhiên. Để ước lượng θ lẽ đương nhiên là phải dựa vào mẫu, cách tốt nhất là ta dùng một hàm nào đó của mẫu, để đơn giản ta ký hiệu hàm đó là $\theta^*(X_1,..., X_n)$.

Như vậy, ta sẽ không xét các ước lượng là hằng số, tức là không phụ thuộc gì vào mẫu.

Ước lượng $\theta^*(X_1,..., X_n)$ là biến ngẫu nhiên, nó sẽ nhận giá trị $\theta^*(x_1,..., x_n)$ khi mẫu nhận giá trị cụ thể $(x_1,..., x_n)$. Giá trị $\theta^*(x_1,..., x_n)$ là một điểm trong miền giá trị của hàm θ^*, cho nên ước lượng $\theta^*(X_1,..., X_n)$ được gọi là ước lượng điểm.

Vì có rất nhiều hàm của mẫu, do đó có rất nhiều ước lượng cho θ, vấn đề là chọn cái nào. Vì vậy, cần xây dựng các tiêu chuẩn để đánh giá các ước lượng.

Ước lượng $\theta^*(X_1,..., X_n)$ được gọi là ước lượng không chệch cho tham ẩn θ nếu thỏa mãn: $E\theta^*(X_1,..., X_n) = \theta$

(Nghĩa là đòi hỏi trung bình của θ^* bằng giá trị cần tìm, hay về trung bình ước lượng θ^* cho ta giá trị θ cần tìm).

Nếu $E\theta^*(X_1,..., X_n) = \theta + b(\theta) \neq \theta$ thì $\theta^*(X_1,..., X_n)$ được gọi là ước lượng chệch, $b(\theta)$ là độ chệch tương ứng.

Ta có các câu trả lời sau:

– Ước lượng điểm cho EX (hoặc cho μ) là $\bar{X}$. Đó là ước lượng không chệch ($E\bar{X} = \mu$).

– Ước lượng điểm cho DX (hoặc cho σ^2) là s^2 hoặc $\hat{s}^2$.

$\hat{s}^2$ là ước lượng không chệch ($E\hat{s}^2 = \sigma^2$), còn s^2 là ước lượng chệch với độ chệch $-\frac{\sigma^2}{n}$; ($Es^2 = \sigma^2 - \frac{\sigma^2}{n}$).

– Ước lượng điểm cho $P(X \in A)$ (hoặc cho p) là: $p^* = \frac{m}{n}$, trong đó n là số quan sát, m là số lần xảy ra biến cố A trong mẫu đã cho. p^* cũng là không chệch ($Ep^* = E(\frac{m}{n}) = \frac{1}{n} E(m) = \frac{np}{n} = p$ (do m là số lần xảy ra biến cố A trong n quan sát độc lập, cho nên m chính là giá trị của biến ngẫu nhiên nhị thức). (Nếu không đọc chương I thì bạn đọc phải công nhận kết quả này).

Trong Thống kê Toán học, người ta đã xây dựng gần chục tiêu chuẩn để đánh giá các ước lượng. Với mức độ của giáo trình này chúng ta chỉ biết được tiêu chuẩn không chệch. Đó là tiêu chuẩn khá đơn giản. Còn các tiêu chuẩn khác, các yêu cầu khác của ước lượng chúng ta không biết được. Vì vậy xin khẳng định với bạn đọc là: Ba ước lượng điểm nói trên là các ước lượng tốt (thậm chí là tốt nhất) theo mọi tiêu chuẩn được xây dựng trong Thống kê Toán học. Vì vậy ở góc độ ứng dụng chúng ta yên tâm sử dụng chúng. Nếu bạn thấy băn khoăn, nghi ngờ về kết luận đưa ra, thì trước tiên hãy kiểm tra lại tính trung thực và khách quan của mẫu đại diện mà bạn đã dùng.

a) Ước lượng điểm cho Mode

Ký hiệu M_0 là ước lượng điểm cho ModX

* Đối với mẫu thu gọn M_0 là giá trị mẫu mà ứng với nó tần số m_i (hoặc tần suất $\frac{m_i}{n}$) là lớn nhất.

Đối với mẫu dạng khoảng: $M_0 = \frac{d_1}{d_1 + d_2} h + L_1$,

trong đó:

Khoảng Mode là khoảng có tần số lớn nhất. Khoảng trước và khoảng sau Mode là hai khoảng trước và sau đối với khoảng Mode.

d_1 – Sự sai khác tuyệt đối giữa tần số của khoảng Mode và khoảng trước Mode.

d_2 – Sự sai khác tuyệt đối giữa tần số của khoảng Mode và khoảng sau Mode.

h – Độ dài của khoảng Mode.

L_1 – Mút trái của khoảng Mode.

Áp dụng công thức trên với số liệu mẫu ở ví dụ II.3 ta có:

Đối với mẫu thu gọn: $M_0 = 750$ và 800.

Đối với mẫu dạng khoảng (4 khoảng):

$$M_0 = \frac{5}{5+3} \cdot 50 + 750 = 781,25$$

b) Ước lượng điểm cho Median (hay trung vị) là Median mẫu, ký hiệu là Med.

Med là giá trị mà nó chia dãy số liệu mẫu đã được sắp thứ tự thành hai phần sao cho số các phần tử mẫu ở trên Med bằng số phần tử mẫu ở dưới Med.

Cách tìm Med:

* Đối với mẫu thu gọn (x_i, m_i) ; $i = \overline{1,k}$:

Giả sử l là vị trí mà $m_1 + ... + m_{l-1} < \frac{n}{2}$, nhưng $m_1 + ... + m_{l-1} + m_l > \frac{n}{2}$

Khi đó Med = x_l

Ở ví dụ II.3 thì:

$l = 5$ vì $6 + 7 + 5 + 5 = 23 < 25 < 6 + 7 + 5 + 5 + 9$

Do đó Med = $x_5 = 750$

Nếu xảy ra trường hợp:

$\sum_{i=1}^{l-1} m_i = \frac{n}{2}$, do đó $m_l + ... + m_k = \frac{n}{2}$ thì Med = $\frac{1}{2}(x_{l-1} + x_l)$

* Đối với mẫu thu gọn dạng khoảng:

Giả sử Med thuộc khoảng thứ l: Med $\in (x_l, x_{l+1})$, tức là:

$$\sum_{i=1}^{l-1} m_i < \frac{n}{2} < \sum_{i=1}^{l} m_i$$

Khi đó: Med = $\frac{\frac{n}{2} - \sum_{i=1}^{l-1} m_i}{m_l}(x_{l+1} - x_l) + x_l$

Trở lại số liệu của ví dụ II.3: n = 50, số khoảng k = 4, khoảng l chính là khoảng 3 vì

$13 + 10 = 23 < 25 < 38 = 13 + 10 + 15$; $x_3 = 750$; $x_4 = 800$.

Do đó: Med = $\frac{25 - 23}{15}.(800 - 750) + 750 = 756,67$

c) Ước lượng điểm cho tứ phân vị

(Ở mức độ 30 tiết thì chúng ta bỏ qua phần này).

Có ba tứ phân vị $x_{\frac{1}{4}}, x_{\frac{2}{4}}, x_{\frac{3}{4}}$ (nhưng $x_{\frac{2}{4}}$ trùng với trung vị).

Ước lượng điểm cho tứ phân vị ta dùng tứ phân vị mẫu, ký hiệu là Q_1, Q_2, Q_3 (Q là viết tắt của từ quartile).

Cách xác định Q_1, Q_3:

* Đối với mẫu không gộp khoảng:

Giả sử mẫu được sắp theo thứ tự $x_{(1)} \le x_{(2)} \le .... \le x_{(n)}$

Nếu n chia hết cho 4 thì Q_1 sẽ là giá trị giữa giá trị thứ $\frac{n}{4}$ và giá trị tiếp theo thứ $\left(\frac{n}{4}+1\right)$: $Q_1 = \frac{1}{2}\left(x_{(\frac{n}{4})} + x_{(\frac{n}{4}+1)}\right)$.

Nếu n không chia hết cho 4 thì Q_1 là giá trị thứ $\left(\frac{n}{4}+1\right)$, tức là giá trị ứng với số nguyên lớn hơn tiếp theo $\frac{n}{4}$: $Q_1 = x_{\left(\frac{n}{4}+1\right)}$

Bằng cách thay $\frac{n}{4}$ bởi $\frac{3n}{4}$, làm tương tự như trên ta xác định được Q_3. Nếu số liệu ở dạng mẫu thu gọn, bạn đọc tự suy ra theo cách trên hoặc làm tương tự như tìm Med (thay $\frac{n}{2}$ bởi $\frac{n}{4}$ hoặc $\frac{3n}{4}$).

Trở lại với ví dụ II.3: $Q_1 = 680$; $Q_2 = Med = 750$; $Q_3 = 780$.

* Đối với mẫu dạng khoảng:

Q_1, Q_3 cũng được xác định như Med (thay $\frac{n}{2}$ bởi $\frac{n}{4}$ và $\frac{3n}{4}$ tương ứng), nghĩa là:

$$Q_s = \frac{s.\frac{n}{4} - \sum_{i=1}^{l-1} m_i}{m_l} .(x_{l+1} - x_l) + x_l \; ; \; s = 1, 2, 3.$$

trong đó l là khoảng thứ l: (x_l, x_{l+1}) mà Q_s rơi vào.

Với ví dụ II.3 ta có: $Q_1 \in [650 ; 700)$; $l = 1$; $m_1 = 13$.

$$Q_1 = \frac{1.\frac{50}{4} - 0}{13} (700 - 650) + 650 = 698,08$$

$Q_2 = Med = 756,67$

$Q_3 \in [750 ; 800)$; $l = 3$; $m_3 = 15$

$$Q_3 = \frac{3.\frac{50}{4} - 23}{15} (800 - 750) + 750 = 798,33$$

Ví dụ II.4: Trở lại ví dụ II.3 với số liệu mẫu đã có:

a) Hãy ước lượng mức thu nhập trung bình của công nhân ở khu chế xuất.

b) Hãy ước lượng bình phương độ tản mát của mức thu nhập của công nhân ở khu chế xuất.

c) Hãy ước lượng độ lệch tiêu chuẩn của mức thu nhập của công nhân ở khu chế xuất.

d) Hãy ước lượng tỷ lệ công nhân có thu nhập thấp (≤ 700.000đ), nghĩa là tỷ lệ công nhân có thu nhập chỉ đủ nuôi sống bản thân ở mức thấp.

Giải:

Như trên ta đã tính được:

$\bar{X} = 737,600 \sim 737600$ đồng ; $s = 55,011272$;

$\hat{s} = 55,569776$. Vậy:

a) Mức thu nhập trung bình của công nhân ở khu chế xuất là 737600 đồng.

b) Bình phương độ tản mát của mức thu nhập là $(55,569776)^2 = 3088,00$

c) Độ lệch tiêu chuẩn của mức thu nhập là 55,569776 đồng.

d) Tỷ lệ công nhân có thu nhập thấp là: $\frac{18}{50} = 0,39 = 39\%$.

Lưu ý rằng bốn câu trả lời trên chỉ là ước lượng, nhưng đó là các ước lượng tốt nhất có thể được. Muốn có câu trả lời đúng với các giá trị thực cần tìm thì ta phải tổng kết toàn bộ số công nhân của cả khu chế xuất, chứ không phải chỉ điều tra ở 50 công nhân đại diện.

Ví dụ II.5: Trong đợt vận động bầu cử Tổng thống, hãng tin nọ đã phỏng vấn ngẫu nhiên 3000 cử tri. Kết quả nhận được 1200 cử tri sẽ bầu cho ứng cử viên A, 1000 cử tri sẽ bầu cho ứng cử viên B, 500 cử tri sẽ ủng hộ ứng cử viên C. Còn 300 cử tri sẽ ủng hộ cho ứng cử viên D. Hãy chỉ ra ước lượng điểm cho tỷ lệ phiếu bầu thực mà ứng cử viên sẽ thu được nếu tổ chức bầu cử ở thời điểm hiện tại.

Giải:

Gọi p_A, p_B, p_C, p_D là tỷ lệ phiếu bầu thực mà các ứng cử viên sẽ nhận được nếu tổ chức bầu cử ở thời điểm hiện tại. Ta nhận được các ước lượng tương ứng như sau:

$$p_A^* = \frac{1200}{3000} = 0,40 = 40\%; \quad p_B^* = \frac{1000}{3000} = 0,333 = 33,3\%;$$

$$p_C^* = \frac{500}{3000} = 0,167 = 16,7\%; \quad p_D^* = \frac{300}{3000} = 0,10 = 10\%.$$

Từ các thông tin trên có thể nói: ở thời điểm hiện tại chưa có ai vượt quá 50%, tức là thắng cử ngay vòng đầu. Ứng cử viên A và B đang dẫn đầu. Ứng cử viên A có khả năng thu được nhiều phiếu ủng hộ nhất. Còn hai ứng cử viên C và D ít khả năng lọt vào vòng sau.

II.3.2. Ước lượng khoảng cho kỳ vọng và xác suất (cho giá trị trung bình và tỷ lệ)

II.3.2.1. Định nghĩa

Một khoảng với hai đầu mút ngẫu nhiên $(\theta_1^*(X_1,..., X_n); \theta_2^*(X_1,..., X_n))$ được gọi là ước lượng khoảng (khoảng tin cậy) cho tham ẩn θ với độ tin cậy $(1 - \alpha)$ nếu thỏa mãn:

$$P\{\theta_1^*(X_1,..., X_n) < \theta < \theta_2^*(X_1,..., X_n)\} \geq 1 - \alpha.$$

Nghĩa là: với xác suất $\geq 1- \alpha$ ta kết luận tham ẩn θ cần tìm sẽ thuộc khoảng $(\theta_1^*; \theta_2^*)$. Kết luận như vậy, khả năng đúng $\geq 1 - \alpha$, nhưng sẽ mắc phải sai lầm với khả năng $\leq \alpha$.

Rõ ràng với cùng một độ tin cậy, khoảng $(\theta_1^*; \theta_2^*)$ càng hẹp thì càng tốt, khoảng càng hẹp thì thông tin càng giá trị, khoảng càng rộng thì thông tin càng ít giá trị.

II.3.2.2. Ước lượng khoảng cho giá trị trung bình

Giả sử $(X_1,..., X_n)$ là mẫu ngẫu nhiên được rút ra từ biến ngẫu nhiên X với $EX = \mu$ chưa biết, đang cần tìm ước lượng khoảng cho μ; còn $DX = \sigma^2$.

a) Trường hợp phương sai $DX = \sigma^2$ đã biết

Khi đó, với giả thiết X có phân phối chuẩn $(X = N(\mu,\sigma^2))$ hoặc cỡ mẫu n đủ lớn $(n \geq 30)$ ta có:

Với độ tin cậy $(1 - \alpha)$ ta kết luận:

$$\mu \in \left(\bar{X} - u\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} ; \bar{X} + u\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} \right).$$

Chứng minh: Từ giả thiết X phân phối chuẩn hoặc cỡ mẫu n lớn ta suy ra $\bar{X} = N\left(\mu; \frac{\sigma^2}{n}\right)$ (xem II.2.4). Với độ tin cậy $1 - \alpha$ đã cho, tra bảng I ta được $u\left(\frac{\alpha}{2}\right)$ (xem phân phối chuẩn công thức (I.10) ở chương I hoặc công thức (4) phần phụ lục I):

$$P\left\{\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} < u\left(\frac{\alpha}{2}\right)\right\} = 1 - \alpha.$$

$$P\left\{\bar{X} - u\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + u\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha.$$

Theo định nghĩa, khoảng $\left(\bar{X} - u\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} ; \bar{X} + u\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} \right)$ là khoảng tin cậy cho μ với độ tin cậy $(1 - \alpha)$. Đó là điều cần chứng minh.

b) Trường hợp DX chưa biết. Giả thiết X có phân phối chuẩn

Khi đó, với độ tin cậy $(1 - \alpha)$ ta kết luận:

$$\mu \in \left(\bar{X} - t_{n-1}\left(\frac{\alpha}{2}\right)\frac{\hat{s}}{\sqrt{n}} ; \bar{X} + t_{n-1}\left(\frac{\alpha}{2}\right)\frac{\hat{s}}{\sqrt{n}} \right).$$

Chứng minh: Vì X chuẩn nên $t = \frac{\bar{X} - \mu}{\hat{s}}\sqrt{n}$ có phân phối Student với tham số $(n - 1)$ (xem II.2.4). Với độ tin cậy $(1 - \alpha)$ đã cho, tra bảng III ta tìm được $t_{n-1}\left(\frac{\alpha}{2}\right)$ sao cho (xem I.9.5 công thức (I.11) hoặc công thức (7) phụ lục I):

$$P\left\{-t_{n-1}\left(\frac{\alpha}{2}\right) < \frac{\bar{X}-\mu}{\hat{s}}\sqrt{n} < t_{n-1}\left(\frac{\alpha}{2}\right)\right\} = 1-\alpha$$

$$P\left\{\bar{X} - t_{n-1}\left(\frac{\alpha}{2}\right)\frac{\hat{s}}{\sqrt{n}} < \mu < \bar{X} + t_{n-1}\left(\frac{\alpha}{2}\right)\frac{\hat{s}}{\sqrt{n}}\right\} = 1-\alpha$$

Đó là điều cần chứng minh. (Nếu dùng s thì lưu ý là $\frac{s}{\sqrt{n-1}} = \frac{\hat{s}}{\sqrt{n}}$)

c) Trường hợp DX chưa biết, giả thiết chuẩn cũng không có nhưng cỡ mẫu n đủ lớn

Khi đó, ta thay σ^2 chưa biết bởi ước lượng điểm không chệch $\hat{s}^2$. Coi như σ^2 đã biết và quay về áp dụng trường hợp a), ta nhận được khoảng xấp xỉ sau:

Với độ tin cậy $(1-\alpha)$ ta kết luận

$$\mu \in \left(\bar{X} - u\left(\frac{\alpha}{2}\right)\frac{\hat{s}}{\sqrt{n}} ; \bar{X} + u\left(\frac{\alpha}{2}\right)\frac{\hat{s}}{\sqrt{n}}\right)$$

Nhận xét:

– Như vậy, giá trị trung bình có 3 ước lượng khoảng. Khoảng c) (khoảng ứng với trường hợp c)) là khoảng xấp xỉ, cho nên tuỳ điều kiện bài toán ta ưu tiên dùng khoảng a) và b).

– Để chọn xem cần dùng khoảng nào (trường hợp nào) ta có hai mốc rẽ nhánh. Mốc rẽ đầu tiên là DX (hoặc σ^2) đã biết hay chưa biết. Mốc rẽ thứ hai là có giả thiết chuẩn hay không. Tất cả các thông tin trên do đề bài cung cấp chứ chúng ta không phải đi tìm hay tính toán.

– Giả thiết n đủ lớn được hiểu là $n \geq 30$. Nhưng đó nên coi là giá trị tối thiểu, n càng lớn càng tốt.

– Các bước làm như sau:

+ Xem đó là ước lượng điểm hay khoảng;

+ Xét xem cần dùng khoảng nào (trường hợp nào);

+ Cho mẫu → Tính $\bar{X}$ hoặc $\bar{X}$ và $\hat{s}$ (tùy trường hợp);

+ Cho độ tin cậy $(1 - \alpha) \rightarrow \alpha \rightarrow \frac{\alpha}{2} \rightarrow$ Tra bảng tìm $u\left(\frac{\alpha}{2}\right)$ hoặc $t_{n-1}\left(\frac{\alpha}{2}\right)$.

+ Thay tất cả vào công thức tính.

Ví dụ II.6: Để nghiên cứu tuổi thọ của một dân tộc thiểu số, người ta thống kê tuổi thọ của những người đã mất của dân tộc đó trong một năm qua ở các vùng miền khác nhau trên cả nước có dân tộc này sinh sống. Kết quả như sau:

Tuổi thọ X (năm)	≤ 3	(3;10]	(10;20]	(20;30]	(30;40]	(40;50]	(50;60]	(60;70]	>70
Số người	15	8	4	3	2	5	20	18	5

a) Hãy ước lượng tuổi thọ trung bình của dân tộc này.

b) Với độ tin cậy 95% tuổi thọ trung bình của dân tộc này thuộc khoảng nào?

c) Với xác suất 90% có thể nói tuổi thọ trung bình của dân tộc này cao nhất là bao nhiêu tuổi?

d) Nếu chỉ xét trong tập những người đã chết trong năm qua của dân tộc này thì các câu hỏi trên sẽ được thay bởi câu hỏi nào?

Giải:

Ở đây ta có n = 80 quan sát. DX chưa biết, không có giả thiết X chuẩn. Từ đó ta có sự nhận biết câu trả lời như sau:

Câu a) là ước lượng điểm cho EX, vì câu hỏi này không cho độ tin cậy.

Câu b) là ước lượng khoảng cho EX. Ta dùng trường hợp c).

Để trả lời câu c) ta dùng đầu mút phải của ước lượng khoảng. (Để đơn giản và hợp lý hơn, ta dùng khoảng đối xứng).

Từ số liệu đã cho ta tính được (vì dùng trường hợp c) nên phải tính cả $\bar{X}$ và $\hat{s}$):

$\bar{X} = 39{,}36875 \approx 39{,}37$; $s = 26{,}552983$; $\hat{s} = 26{,}720511$.

(Khoảng số liệu đầu: ≤ 3, được hiểu là [0; 3], điểm giữa là 1,5; khoảng cuối > 70 được hiểu là (70; +∞), nhưng một cách thực tế hơn ta hiểu mút cuối của khoảng là 80 hoặc 90, hoặc 100,... Điểm đại diện ta lấy là 75 để cho cách đều với các điểm đại diện ở trước: 25, 35, 45, 55, 65).

a) Tuổi thọ trung bình của dân tộc thiểu số này là 39,37 tuổi.

b) Dùng khoảng c), tra bảng ta được u(0,025) = 1,96.

Với độ tin cậy 95% ta kết luận

$$EX \in (39{,}37 \pm 1{,}96.\frac{26{,}721}{\sqrt{80}}) \leftrightarrow EX \in (33{,}51 \text{ tuổi}; 45{,}23 \text{ tuổi}).$$

c) Ta vẫn dùng khoảng c). Với xác suất 90% có thể nói tuổi thọ trung bình của dân tộc thiểu số này cao nhất là:

$$39{,}75 + 1{,}65.\frac{26{,}721}{\sqrt{80}} = 44{,}68 \text{ tuổi } (u(0{,}05) = 1{,}65).$$

d) Nếu kết luận cho cả dân tộc thiểu số đang xét thì tuổi thọ của 80 người trong năm qua chỉ là một mẫu đại diện, nhưng nếu chỉ xét trong tập những người đã chết trong năm qua của dân tộc này thì tập số liệu trên là số liệu được thống kê một cách đầy đủ. Vì vậy, để trả lời cho tuổi thọ trung bình của những người đã mất thì ba câu hỏi trên sẽ được thay bởi câu hỏi: Hãy tính tuổi thọ trung bình của những người đã mất trong một năm qua của dân tộc thiểu số đang xét.

Ví dụ II.7: Để xác định chiều cao của các em lứa tuổi lên 10 ở nông thôn vùng đồng bằng Bắc Bộ người ta lấy ra một mẫu đại diện với các kết quả như sau:

Khoảng chiều cao X (cm)	< 130	[130;135)	[135;140)	[140;145)	≥145
Số em	5	15	30	20	5

Giả sử chiều cao X tuân theo luật phân phối chuẩn với DX = 9.

a) Hãy ước lượng chiều cao trung bình của các em lứa tuổi lên 10 ở nông thôn vùng đồng bằng Bắc Bộ. Hãy ước lượng cho ModX.

b) Với độ tin cậy 90% thì có thể kết luận chiều cao trung bình của các em thuộc khoảng nào? Khả năng đúng của kết luận là bao nhiêu? Khả năng sai là bao nhiêu?

c) Với xác suất 96% có thể nói chiều cao trung bình của các em thấp nhất là bao nhiêu cm?

Giải: Theo đề bài: DX = 9 đã biết, X có phân phối chuẩn. Từ số liệu mẫu đã cho ta tính được $\overline{X}$ = 137,83 (Vì $\sigma^2 = 9$ đã biết, nên ta không cần tính s hay ŝ nữa).

a) Đây là ước lượng điểm cho EX nên ta dùng $\overline{X}$ để trả lời.

Chiều cao trung bình của các em lứa tuổi lên 10 ở nông thôn đồng bằng Bắc Bộ là 137,83 cm.

Ước lượng điểm cho ModX là: $M_0 = \frac{15}{15+10} \cdot 5 + 135 = 138$.

b) $\sigma^2 = 9$ đã biết, X có phân phối chuẩn. Ta dùng khoảng a) để trả lời. Tra bảng ta có u(0,05) = 1,65.

$$EX \in (137,83 \pm 1,65.\frac{3}{\sqrt{75}}) = (137,26m;\ 138,40cm).$$

Trung bình thực EX chưa biết. Nhưng ta kết luận với khả năng 90% EX sẽ thuộc khoảng (137,26cm; 138,40cm). Kết luận này sẽ có khả năng đúng là 90% nhưng cũng mắc khả năng sai là 10%.

c) Với xác suất 0,96 có thể nói chiều cao trung bình thấp nhất là:

$$137,83 - 2,06.\frac{3}{\sqrt{75}} = 137,12\ cm;\ (u(0,02) = 2,06)$$

II.3.2.3. Ước lượng khoảng cho tỷ lệ

Gọi $p = P(X \in A)$ hoặc p là tỷ lệ nào đó. Giả thiết n đủ lớn. Khi đó với độ tin cậy $(1 - \alpha)$ ta kết luận:

$$p \in \left(p^* - u(\frac{\alpha}{2})\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}}\ ;\ p^* + u(\frac{\alpha}{2})\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}} \right)$$

Chứng minh: Ở mức độ 30 tiết thì bỏ qua phần chứng minh này. Ở mức độ 45 tiết ta có thể chứng minh như sau: Liên quan đến xác suất p, chúng ta xét một mẫu ngẫu nhiên ứng với n phép thử Bernoulli (xem I.5.1): $P(X_i = 1) = p$; $P(X_i = 0) = 1 - p$. Khi đó $EX_i = p$, $DX_i = p(1 - p)$. Tham ẩn p lại đóng vai trò của kỳ vọng μ, do đó để tìm ước lượng khoảng cho p ta dùng trường hợp c) của ước lượng khoảng cho giá trị trung bình μ. Ta có $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{m}{n} = p^*$; $\sigma^2 = p(1 - p)$ chưa biết nên ta thay bởi ước lượng $p^*(1-p^*)$.

Theo trường hợp c) ta được khoảng tin cậy cho p là:

$$\left(p^* - u(\frac{\alpha}{2})\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}}\ ;\ p^* + u(\frac{\alpha}{2})\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}} \right)$$

Ví dụ II.8: Trở lại số liệu ở ví dụ II.6.

a) Hãy ước lượng tỷ lệ người có tuổi thọ trên 60 tuổi của dân tộc này. Với xác suất 98% tỷ lệ trên thấp nhất là bao nhiêu %?

b) Hãy ước lượng tỷ lệ trẻ em ≤ 3 tuổi bị chết (trong số những người bị chết). Với xác suất 98% tỷ lệ này thuộc khoảng nào? Cao nhất là bao nhiêu?

c) Cho biết trong một năm qua số trẻ được sinh ra của dân tộc này là 120. Trong số 15 trẻ bị chết dưới 3 tuổi thì có 8 trường hợp là trẻ sơ sinh. Hãy tính tỷ lệ trẻ sơ sinh bị chết. Tỷ lệ này là giá trị thực hay ước lượng. Trong tình huống nào tỷ lệ này chỉ được coi là ước lượng?

Giải:

a) Gọi P là tỷ lệ người thọ trên 60 tuổi của dân tộc này hay p = P(X > 60).

Ước lượng điểm cho p là $p^* = \frac{18+5}{80} = 0,2875 = 28,75\%$.

Với xác suất 98% tỷ lệ người thọ trên 60 tuổi của dân tộc này thấp nhất là:

$$0,2875 - 2,33.\frac{\sqrt{0,2875 \,.\, 0,7125}}{\sqrt{80}} = 0,2764 = 27,64\%.$$

(Tra bảng ta có u(0,01) = 2,33).

b) Đặt p = P(X ≤ 3) = Tỷ lệ trẻ em ≤ 3 tuổi bị chết.

Ta có ước lượng điểm $p^* = \frac{15}{80} = 0,1875 = 18,75\%$.

Với xác suất 98%:

$$p \in (0,1875 \pm 2,33.\frac{\sqrt{0,1875.0,8125}}{\sqrt{80}}) = (0,1773;\ 0,1977).$$

Với xác suất 98% tỷ lệ trẻ em ≤ 3 tuổi của dân tộc này bị chết cao nhất là 0,1977 = 19,77%.

c) Tỷ lệ trẻ sơ sinh bị chết là: $\frac{8}{120} = 0,0667 = 6,67\%$.

Theo bài toán, 80 người chết và 120 trẻ được sinh ra trong một năm qua của dân tộc này là số liệu điều tra đầy đủ. Vì vậy, tỷ lệ 6,67% ở trên là tỷ lệ thực trong năm đang xét. Nhưng nếu coi hai con số 80 và 120 của năm này là một mẫu đại diện cho nhiều năm của dân tộc này, khi đó 6,67% lại là giá trị ước lượng, nghĩa là tỷ lệ trẻ sơ sinh của dân tộc này bị chết được ước lượng là 6,67%.

Ví dụ II.9: Trở lại ví dụ II.5.

a) Với độ tin cậy 0,95 tỷ lệ phiếu bầu của ứng cử viên A và B thuộc khoảng nào? Cao nhất là bao nhiêu %? Khả năng đúng, sai là bao nhiêu?

b) Với xác suất 90% có thể nói tỷ lệ phiếu bầu của ứng cử viên A thấp nhất là bao nhiêu %, còn của ứng cử viên B cao nhất là bao nhiêu %?

Giải:

Tra bảng ta được u(0,025) = 1,96; u(0,05) = 1,65.

a) Với xác suất 0,95 ta kết luận:

$$p_A \in (0,40 \pm 1,96 . \frac{\sqrt{0,4.0,6}}{\sqrt{3000}}) = (0,382 ; 0,418).$$

$$p_B \in (0,333 \pm 1,96 . \frac{\sqrt{0,333.0,667}}{\sqrt{3000}}) = (0,319 ; 0,347).$$

Với xác suất 0,95 tỷ lệ p_A cao nhất là 41,8%, còn p_B cao nhất là 34,7%. Các kết luận trên có khả năng đúng là 95%, nhưng khả năng sai là 5%.

b) Với xác suất 90% tỷ lệ phiếu bầu p_A của ứng cử viên A thấp nhất là:

$$0,40 - 1,65 \cdot \frac{\sqrt{0,4 \cdot 0,6}}{\sqrt{3000}} = 0,385 = 38,5\%.$$

Trong khi đó, với xác suất 90% tỷ lệ phiếu bầu p_B của ứng cử viên B cao nhất là:

$$0,333 + 1,65 \cdot \frac{\sqrt{0,333 \cdot 0,667}}{\sqrt{3000}} = 0,345 = 34,5\%.$$

II.4. MỘT SỐ BÀI TOÁN KIỂM ĐỊNH GIẢ THIẾT ĐƠN GIẢN

II.4.1. Đặt bài toán

Trong thực tế, chúng ta thường gặp các tình huống, các bài toán dạng như sau:

Trong cuộc vận động bầu cử, các cố vấn của ứng cử viên A có những nhận định khác nhau. Một số cho rằng ứng cử viên A sẽ vượt qua 50% số phiếu, tức là ứng cử viên A sẽ thắng cử ngay vòng I. Nhưng lại có một số ý kiến cho rằng ứng cử viên A chưa thể vượt qua vòng I. Hoặc qua thăm dò với một mẫu đại diện người ta nhận được ước lượng cho tỷ lệ phiếu bầu của ứng cử viên A là 42%, của ứng cử viên B là 39%. Liệu có thể khẳng định ứng cử viên A sẽ thu được nhiều phiếu hơn ứng cử viên B khi bầu cử thật hay không?

Có ý kiến cho rằng học sinh trường THCS A học môn Toán tốt hơn học sinh trường THCS B, song lại có ý kiến chưa nhất trí.

Có ý kiến cho rằng tình trạng đạo đức của thanh thiếu niên phụ thuộc vào hoàn cảnh gia đình, nhưng lại có ý kiến chưa tán thành v.v...

Vậy ý kiến nào là có cơ sở, ý kiến nào được chấp nhận?

Bài toán đặt ra là khi có hai tình huống (hai giả thiết) về một vấn đề nào đó, yêu cầu chúng ta phải chọn lấy một và loại bỏ một trên cơ sở thông tin từ một mẫu đại diện. Dĩ nhiên, yêu cầu là chọn tình huống nào để khả năng đúng cao hơn, khả năng sai thấp hơn. Yêu cầu như vậy là hợp lý và luôn có câu trả lời. Tất nhiên, không thể yêu cầu cho một câu trả lời đúng hoàn toàn, không mắc sai lầm nào, khi mà thông tin của chúng ta chỉ là một mẫu đại diện, vừa không đầy đủ, vừa chứa sai số.

Vậy, với yêu cầu như trên, trong hai tình huống đưa ra đều có khả năng sai thì chúng ta vẫn cứ phải chọn lấy một và hãy chọn tình huống nào có khả năng sai ít hơn.

Thực ra đây là bài toán hãy chọn một quyết định khi mà thông tin duy nhất ta có là thông tin không những không đầy đủ mà còn chứa thông tin nhiễu. Đó là bài toán hay gặp trên thực tế trong tự nhiên, xã hội và trong cuộc đời của mỗi người. Thật là khó khăn khi phải chọn một quyết định dựa trên thông tin không đầy đủ. Đòi hỏi chọn một quyết định hoàn toàn đúng là đòi hỏi không thực tế. Hãy cố gắng chọn quyết định nào để khả năng đúng cao nhất, khả năng sai thấp nhất.

Để cho tiện, một trong hai tình huống đưa ra ta gọi là giả

thiết H, còn tình huống kia gọi là đối thiết K.

Ý cơ bản để giải bài toán này là: Chia miền giá trị có thể của mẫu ngẫu nhiên thành hai phần S và $\overline{S}$ (tất nhiên không phải chia tùy ý, nhưng chia thế nào, ta sẽ đề cập tới sau) và thực hiện theo quy tắc sau:

– Nếu mẫu $(X_1,..., X_n) \in S$ thì ta bác bỏ H (do đó tạm chấp nhận K)

– Nếu mẫu $(X_1,..., X_n) \in \overline{S}$ thì chưa có cơ sở bác bỏ H (do đó tạm chấp nhận H và bác bỏ K).

Quyết định như vậy cho đến khi nào có thông tin mới.

Làm theo quy tắc trên sẽ phạm phải hai sai lầm:

– Sai lầm loại I: Bác bỏ giả thiết H, nhưng thực tế H đúng.

– Sai lầm loại II: Chấp nhận giả thiết H, nhưng thực tế H sai.

Tất nhiên ta mong muốn chọn giả thiết nào để cực tiểu cả hai khả năng phạm sai lầm. Nhưng khi cỡ mẫu n cố định, trong Thống kê Toán học đã chỉ ra rằng: bài toán không có lời giải. Vì vậy, người ta khống chế sai lầm loại I và cực tiểu khả năng phạm sai lầm loại II, nghĩa là người ta cho trước một giới hạn trên α của xác suất phạm sai lầm loại I (α thường nhỏ) và bài toán đưa đến là: Hãy chọn giả thiết nào (tức là chọn S = ?) để sao cho:

1°. Xác suất phạm sai lầm loại I không vượt quá α.

2°. Khả năng phạm sai lầm loại II đạt cực tiểu.

Số α được gọi là mức ý nghĩa của bài toán.

Miền S là miền bác bỏ H, được gọi là miền tiêu chuẩn.

Như vậy, khi phát biểu bài toán kiểm định giả thiết, ta phải chỉ rõ H, K, α.

Câu trả lời cho bài toán chính là miền bác bỏ H hay miền S là miền nào. (Lưu ý rằng miền S phải thỏa mãn hai yêu cầu 1^*, 2^* ở trên. Tại sao tìm được miền S như thế, việc chứng minh miền S phải thỏa mãn 1^* và 2^* sẽ không đặt ra ở giáo trình này. Đó là công việc của các nhà toán học. Chúng ta chỉ cần hiểu cách đặt bài toán và biết dùng chúng là được).

II.4.2. Kiểm định giả thiết về giá trị trung bình

$$H: \quad \mu = \mu_0 / K \quad (EX = \mu_0) \quad \begin{cases} \mu \neq \mu_0 \\ \mu > \mu_0 ; \alpha \\ \mu < \mu_0 \end{cases}$$

Giá trị trung bình μ chưa biết, μ_0 là một giá trị đã cho. Viết như trên thực chất là ta có ba bài toán kiểm định giả thiết về giá trị μ cần tìm. Câu trả lời (miền bác bỏ giả thiết S) đối với ba bài toán tương đối giống nhau, cho nên để tránh lặp lại, chúng ta xét cùng lúc cả ba bài toán. (Cũng qua đó bạn đọc rút ra các nhận xét về sự giống nhau, khác nhau cũng như quy luật của ba miền tiêu chuẩn tương ứng với ba bài toán dạng trên. Có được các nhận xét trên bạn đọc sẽ thấy việc nhớ và suy ra các miền tiêu chuẩn xuất hiện ở mục II.4 này (gần 30 miền tiêu chuẩn) sẽ đơn giản hơn).

Để giải ba bài toán trên, chúng ta lại phân chia thành ba trường hợp giải (Cách phân chia hoàn toàn tương tự như trong ước lượng khoảng cho giá trị trung bình):

a) Nếu $DX = \sigma^2$ đã biết. Giả thiết X có phân phối chuẩn hoặc cỡ mẫu n đủ lớn

Câu trả lời cho ba bài toán trên là ba miền tiêu chuẩn tương ứng như sau:

$$S_1 = \left\{ \frac{\left|\bar{X} - \mu_0\right|}{\frac{\sigma}{\sqrt{n}}} \geq u\left(\frac{\alpha}{2}\right) \right\} \quad \text{(II.1)}$$

$$S_2 = \left\{ \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq u(\alpha) \right\} \quad \text{(II.2)}$$

$$S_3 = \left\{ \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq -u(\alpha) \right\} \quad \text{(II.3)}$$

b) Nếu DX chưa biết. X có phân phối chuẩn

Khi đó ba miền tiêu chuẩn tương ứng là (nếu dùng s thì $\frac{\hat{s}}{\sqrt{n}} = \frac{s}{\sqrt{n-1}}$):

$$S_1 = \left\{ \frac{\left|\bar{X} - \mu_0\right|}{\frac{\hat{s}}{\sqrt{n}}} \geq t_{n-1}\left(\frac{\alpha}{2}\right) \right\} \quad \text{(II.4)}$$

$$S_2 = \left\{ \frac{\bar{X} - \mu_0}{\frac{\hat{s}}{\sqrt{n}}} \geq t_{n-1}(\alpha) \right\} \quad \text{(II.5)}$$

$$S_3 = \left\{ \frac{\bar{X} - \mu_0}{\frac{\hat{s}}{\sqrt{n}}} \leq -t_{n-1}(\alpha) \right\} \quad \text{(II.6)}$$

c) Nếu DX chưa biết, giả thiết X chuẩn không có, nhưng cỡ mẫu n đủ lớn

Khi đó, cũng như ước lượng cho μ, ta sử dụng trường hợp a) bằng cách thay σ bởi $\hat{s}$. Ký hiệu $u = \dfrac{\bar{X} - \mu_0}{\dfrac{\hat{s}}{\sqrt{n}}}$, ta nhận được:

$$S_1 = \left\{ |u| \geq u\left(\frac{\alpha}{2}\right) \right\} \qquad (II.7)$$

$$S_2 = \{u \geq u(\alpha)\} \qquad (II.8)$$

$$S_3 = \{u \leq -u(\alpha)\} \qquad (II.9)$$

Nhận xét:

– Các miền tiêu chuẩn được chỉ ra ở trên thỏa mãn hai yêu cầu 1°, 2° của bài toán kiểm định giả thiết. Với mức độ của giáo trình chúng ta phải công nhận điều đó. (Thực ra với các kết quả về xác suất được trình bày ở chương I chúng ta sẽ chứng minh được yêu cầu 1°). Thay cho việc chứng minh các S_i thỏa mãn 1° và 2°, dưới đây chúng tôi sẽ trình bày ý tưởng đời thường của việc bác bỏ giả thiết:

Trong xác suất, ta đã biết rằng một biến cố xảy ra với xác suất nhỏ vẫn có thể xảy ra. Nhưng trong một số tình huống cụ thể biến cố với xác suất lớn lại không xảy ra mà xảy ra biến cố với xác suất nhỏ là một điều không bình thường. Chẳng hạn, em Z nào đó được mọi người nói là học giỏi. Khi đó, khả năng Z bị trượt tốt nghiệp sẽ nhỏ, chẳng hạn 10%. Như vậy, nếu giả thiết Z học giỏi là đúng thì biến cố Z bị trượt tốt nghiệp xảy ra chỉ với xác suất nhỏ. Nếu biến cố này xảy ra (tức là Z bị trượt tốt nghiệp), đó sẽ là điều không bình thường. Do đó, chúng ta phải nghi ngờ tính đúng đắn của các giả thiết và đành tạm thời bác bỏ nó cho đến khi có thông tin mới. (Bạn có thấy quyết định như vậy là hợp lý hay không?). Quyết định như vậy cũng sẽ có khả năng oan cho Z, nhưng khả năng đó nhỏ, chỉ là 10% (chính là khả năng ta phạm phải sai lầm loại I).

Xét bài toán: H: $\mu = \mu_0$ / K: $\mu > \mu_0$; α ứng với trường hợp a), miền S sẽ là S_2(II.2). Với giả thiết X có phân phối chuẩn, theo II.2.4 ta có

$\overline{X} = N\left(\mu; \frac{\sigma^2}{n}\right)$. Nếu giả thiết $\mu = \mu_0$ đúng thì $\overline{X} \simeq N\left(\mu_0; \frac{\sigma^2}{n}\right)$. Theo kết quả về phân phối chuẩn (xem I.9.3 công thức (I.9) hoặc công thức (2) phụ lục I) ta có:

$$P\left\{\frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq u(\alpha)\right\} = \alpha$$

hay $P(S_2$ / trong tình huống $\mu = \mu_0) = \alpha$

Nghĩa là: xác suất {bác bỏ H: $\mu = \mu_0$, nhưng thực ra H đúng} là α, hay thỏa mãn yêu cầu 1°.

Như vậy, nếu H đúng thì miền S_2(II.2) sẽ xảy ra với xác suất nhỏ bằng α, cho nên nếu thấy miền S_2(II.2) xảy ra thì chúng ta phải nghi ngờ tính đúng đắn của giả thiết H. Bạn đọc hãy hiểu kỹ nhận xét và cách giải thích cho miền S_2 này, để có thể tự giải thích tính hợp lý cho tất cả các miền tiêu chuẩn khác trong mục II.4.

Tương tự như vậy cho các miền S_1(II.1), S_3(II.3) và các trường hợp b), c).

– Các bước thực hiện bài toán kiểm định giả thiết gồm:

* Nhận ra bài toán kiểm định giả thiết (một dấu hiệu rất rõ là bài toán cho mức ý nghĩa α).

* Nhận ra giả thiết H, đối thiết K (vì H đã được xác định duy nhất, cho nên chỉ còn việc nhận ra đối thiết K. Căn cứ vào câu hỏi của bài toán để xem đối thiết nào cần được chọn).

* Nhận ra trường hợp giải (giống như đã làm trong bài toán ước lượng khoảng).

* Cho mẫu $\rightarrow$ tính $\overline{X}$ và $u = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ hoặc tính $\overline{X}$, $\hat{s}$ và $t = \frac{\overline{X} - \mu_0}{\frac{\hat{s}}{\sqrt{n}}}$ (tùy trường hợp).

* Với mức ý nghĩa α đã cho, tra bảng tìm $u\left(\frac{\alpha}{2}\right)$, $u(\alpha)$ hoặc $t_{n-1}\left(\frac{\alpha}{2}\right)$, $t_{n-1}(\alpha)$ (tùy trường hợp).

* So sánh (căn cứ vào miền tiêu chuẩn S của bài toán), rồi rút ra kết luận (dựa vào miền S xảy ra hay $\overline{S}$ xảy ra).

Ví dụ II.10: Trở lại ví dụ II.7.

Từ ước lượng điểm nhận được, có thể kết luận chiều cao trung bình của các em lứa tuổi lên 10 ở nông thôn ĐBBB cao hơn 137cm; cao hơn 137,5cm được không (xét với mức ý nghĩa $\alpha = 0,05$)?

Giải:

Bài toán cho mức ý nghĩa $\alpha = 0,05$. Bài toán quan tâm đến chiều cao trung bình. Do đó, đây là bài toán kiểm định về giá trị trung bình EX. Theo câu hỏi của bài, ta cần xét hai đối thiết EX > 137 và EX > 137,5. Tất nhiên, giả thiết được xác định là dấu bằng: EX = 137 và EX = 137,5. Ở đây DX = 9 đã biết, X chuẩn cho nên ta dùng trường hợp a), miền tiêu chuẩn cần dùng là S_2. Ta tính được $\overline{X} = 137,83$cm.

H: EX = 137 / K: EX > 137 ; $\alpha = 0,05$

$$u = \frac{137,83 - 137}{\frac{3}{\sqrt{75}}} = 2,396 \; ; \; u(0,05) = 1,65$$

$2,396 > 1,65 \rightarrow$ miền S_2(II.2) xảy ra $\rightarrow$ bác bỏ H: EX = 137, chấp nhận EX > 137cm.

H: EX = 137,5 / K: EX > 137,5 ; $\alpha = 0,05$

$$u = \frac{137,83 - 137,5}{\frac{3}{\sqrt{75}}} = 0,953 < 1,65 \text{ miền } \bar{S}_2 \text{ (II.2) xảy ra} \rightarrow$$

chấp nhận EX = 137,5; chưa có cơ sở chấp nhận EX > 137,5 cm.

Như vậy, ước lượng điểm cho EX là 137,83cm, lớn hơn 137 và lớn hơn 137,5 nhưng dùng tiêu chuẩn toán học, với mức ý nghĩa $\alpha = 5\%$ ta kết luận EX bằng 137,5cm và EX > 137cm.

II.4.3. Kiểm định giả thiết về tỷ lệ

Gọi p là $P(X \in A)$ hoặc là một tỷ lệ nào đó (giá trị thực p ta chưa biết).

$$H: p = p_0 \;/\; K: \begin{cases} p \neq p_0 \\ p > p_0 \\ p < p_0 \end{cases} ; \alpha$$

p_0 là một giá trị xác suất đã cho ($0 \leq p_0 \leq 1$).

Giả thiết cần có là n đủ lớn. Từ mẫu đã cho ta tìm được số lần xảy ra biến cố A là m. Từ đó nhận được $p^* = \frac{m}{n}$ và tính được $u = \frac{p^* - p_0}{\sqrt{p_0(1-p_0)}} . \sqrt{n}$. Ba miền tiêu chuẩn tương ứng với ba bài toán trên là:

$$S_1 = \left\{ |u| \geq u\left(\frac{\alpha}{2}\right) \right\} \qquad \text{(II.10)}$$

$$S_2 = \{ u \geq u(\alpha) \} \qquad \text{(II.11)}$$

$$S_3 = \{ u \leq -u(\alpha) \} \qquad \text{(II.12)}$$

(Đến đây chúng ta đã nhận ra sự giống và khác nhau cũng như quy luật của các miền tiêu chuẩn này. Đối với ba miền trên ta chỉ cần nhớ $u = \frac{p^* - p_0}{\sqrt{p_0(1-p_0)}}.\sqrt{n}$. Từ đó ta suy ra các miền S_α. Các bước làm tương tự như phần trước, chỉ khác là không phải nhận ra trường hợp giải (vì chỉ có một trường hợp), việc tính $\overline{X}$ được thay bởi tính p^*).

Ví dụ II.11: Nghiên cứu về đời sống của dân tộc ở các vùng sâu, vùng xa, tiêu chí được khảo sát là có nhà vệ sinh hợp quy cách và có nước mưa hoặc nước giếng để dùng. Điều tra bằng cách mỗi huyện chọn ngẫu nhiên 1/3 số xã, mỗi xã chọn ngẫu nhiên 1/3 số thôn, mỗi thôn chọn ngẫu nhiên 1/3 số hộ gia đình. Ta nhận được mẫu đại diện sau: Số hộ gia đình được điều tra là 180. Số hộ có nhà vệ sinh hợp quy cách là 124, số hộ có nước mưa hoặc nước giếng dùng là 146.

a) Với mức ý nghĩa $\alpha = 0,10$ có thể kết luận tỷ lệ các hộ gia đình có nhà vệ sinh hợp quy cách nhỏ hơn 70% hay không?

b) Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận tỷ lệ các hộ gia đình có nước mưa hoặc nước giếng dùng lớn hơn 80% hay không?

Giải:

Rõ ràng theo đề bài, ta thấy ngay đây là hai bài toán kiểm định về tỷ lệ với đối thiết $K < 0,70$ và $K > 0,80$. Cụ thể:

a) Gọi p là tỷ lệ các hộ gia đình dân tộc ở vùng sâu vùng xa có nhà vệ sinh hợp quy cách.

H: $p = 0,70$ / K: $p < 0,70$; $\alpha = 0,10$.

$$p^* = \frac{124}{180} = 0,689 \rightarrow u = \frac{0,689 - 0,70}{\sqrt{0,70 \,.\, 0,30}}.\sqrt{180} = -0,322$$

$u(0,10) = -1,28$ So sánh: $-0,322 > -1,28$

Miền $\overline{S_3}(2)$ (II.12) xảy ra nên ta chấp nhận p là 70%.

b) Gọi p là tỷ lệ các hộ gia đình dân tộc ở vùng sâu vùng xa có nước mưa hoặc nước giếng dùng. Ta có:

H: $p = 0,80$ / K: $p > 0,80$; $\alpha = 0,05$

$$p^* = \frac{146}{180} = 0,811 \rightarrow u = \frac{0,811 - 0,80}{\sqrt{0,80 \,.\, 0,20}} \,.\, \sqrt{180} = 0,369$$

$u(0,05) = 1,65 \rightarrow 0,369 < 1,65.$

Miền $\overline{S_2}(11)$ (II.11) xảy ra nên ta chấp nhận p là 80%.

II.4.4. So sánh hai giá trị trung bình

Đây là bài toán được dùng khá rộng rãi. Mỗi khi cần so sánh hai tập hoặc hai biến, đặc trưng đầu tiên thường hay được dùng chính là giá trị trung bình. Giả sử ta có hai giá trị trung bình μ_1, μ_2 chưa biết.

$$\text{H: } \mu_1 = \mu_2 \quad / \quad \text{K} \begin{cases} \mu_1 \neq \mu_2 \\ \mu_1 > \mu_2 \\ \mu_1 < \mu_2 \end{cases} ; \alpha$$

$(EX = EY)$

Để giải quyết ba bài toán kiểm định này chúng ta cần có hai mẫu được rút ra từ hai tập hoặc hai biến tương ứng. Hai mẫu này được rút ra một cách độc lập với nhau. (Vì vai trò μ_1, μ_2 là như nhau cho nên bài toán 3: $\mu_1 = \mu_2$ / $\mu_1 < \mu_2$ có thể đưa về bài toán 2: $\mu_1 = \mu_2/\mu_1 > \mu_2$ bằng cách thay đổi tương ứng). Giải quyết ba bài toán này chúng ta có 5 trường hợp giải. Ba trường hợp đầu được phân chia và nhận biết quen thuộc như trong bài toán ước lượng khoảng hay kiểm định giả thiết đối với giá trị trung bình.

a) Nếu $DX = \sigma_1^2$, $DY = \sigma_2^2$ đã biết. Giả thiết X, Y có phân phối chuẩn hoặc cỡ mẫu n_1, n_2 đủ lớn.

Khi đó ba miền tiêu chuẩn tương ứng là:

$$S_1 = \left\{ \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq u\left(\frac{\alpha}{2}\right) \right\} \qquad \text{(II.13)}$$

$$S_2 = \left\{ \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq u(\alpha) \right\} \qquad \text{(II.14)}$$

$$S_3 = \left\{ \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq -u(\alpha) \right\} \qquad \text{(II.15)}$$

b) Nếu DX, DY chưa biết. Giả thiết X, Y có phân phối chuẩn. Hơn nữa đòi hỏi DX = DY.

Khi đó:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{n_1 s_x^2 + n_2 s_y^2}} \cdot \frac{\sqrt{(n_1 + n_2 - 2) n_1 n_2}}{\sqrt{n_1 + n_2}}$$

và

$$S_1 = \left\{ |t| \geq t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) \right\} \qquad \text{(II.16)}$$

$$S_2 = \{ t \geq t_{n_1+n_2-2}(\alpha) \} \qquad \text{(II.17)}$$

$$S_3 = \{ t \leq - t_{n_1+n_2-2}(\alpha) \} \qquad \text{(II.18)}$$

c) Nếu DX, DY chưa biết, X, Y không tuân theo phân phối chuẩn, nhưng cỡ mẫu n_1, n_2 đủ lớn

Quay về trường hợp a) với:

$$u = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{s}_x^2}{n_1} + \frac{\hat{s}_y^2}{n_2}}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_x^2}{n_1 - 1} + \frac{s_y^2}{n_2 - 1}}}$$

$$S_1 = \left\{ |u| \geq u\left(\frac{\alpha}{2}\right) \right\} \qquad \text{(II.19)}$$

$$S_2 = \{ u \geq u(\alpha) \} \qquad \text{(II.20)}$$

$$S_3 = \{ u \leq -u(\alpha) \} \qquad \text{(II.21)}$$

Các miền tiêu chuẩn từ (II.13) đến (II.21) đều thỏa mãn hai yêu cầu 1° và 2° của bài toán kiểm định giả thiết.

Nếu giả thiết H đúng thì các đại lượng u hoặc t ở trên sẽ có phân phối chuẩn hoặc phân phối Student (xem II.2.4c), do đó các miền S_i chỉ xảy ra với xác suất α nhỏ, cho nên thấy S_i xảy ra chúng ta nghi ngờ H, dẫn đến việc bác bỏ H.

Các bước thực hiện hoàn toàn tương tự như kiểm định giá trị trung bình $\mu = \mu_0$, chỉ khác ở khâu: cho hai mẫu ta phải tính $\overline{X}$, $\overline{Y}$ hoặc $\overline{X}$, $\overline{Y}$ và s_x, s_y.

Ví dụ II.12: Để đánh giá kết quả học tập môn Văn – Tiếng Việt của hai trường THCS A và B khối lớp 9, người ta lấy ra hai mẫu đại diện từ kết quả thi kiểm tra chất lượng do Quận tổ chức:

Trường A:	$n_1 = 80$	$\overline{X} = 6,5$	$s_x = 1,0$
Trường B:	$n_2 = 100$	$\overline{Y} = 7,0$	$s_y = 1,2$

Giả sử điểm thi X, Y của hai trường đều có phân phối chuẩn và DX = DY.

a) Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận trường B có kết quả thi môn Văn - Tiếng Việt tốt hơn trường A hay không?

b) Với độ tin cậy 95% có thể nói điểm trung bình trường A cao nhất là bao nhiêu, trường B thấp nhất là bao nhiêu?

Giải:

a) Nhận ra ngay đây là so sánh hai giá trị trung bình với

H: EX = EY / K: EX < EY; $\alpha = 0,05$

Ta phải dùng trường hợp b) để giải. Miền tiêu chuẩn là S_3(II.18).

Vậy:

$$t = \frac{6,5 - 7,0}{\sqrt{80 \,.\, 1 + 100 \,.\, 1,2^2}} \,.\, \frac{\sqrt{(80 + 100 - 2).80.100}}{\sqrt{80 + 100}} = -2,97.$$

Tra bảng ta được $t_{178}(0,05) \approx 1,66$

$-2,97 < -1,66$. Miền S_3(II.18) xảy ra $\rightarrow$

Ta bác bỏ EX = EY và chấp nhận EX < EY, tức là có thể kết luận trường B học môn Văn – Tiếng Việt tốt hơn trường A.

b) Câu hỏi này liên quan tới đầu mút của ước lượng khoảng cho trung bình. Ta dùng khoảng b) vì DX, DY chưa biết; X, Y có phân phối chuẩn.

Tra bảng ta được $t_{79}(0,025) \approx 1,99$; $t_{99}(0,025) \approx 1,98$

Với độ tin cậy 95% điểm trung bình trường A cao nhất là:

$$6,5 + 1,99.\frac{1}{\sqrt{79}} \approx 6,724;$$

Còn điểm trung bình trường B thấp nhất là:

$$7 - 1,98.\frac{1,2}{\sqrt{99}} \approx 6,761.$$

Ví dụ II.13: Nghiên cứu trọng lượng của trẻ em lứa tuổi lên 10 ở thành phố và nông thôn, ta có hai mẫu đại diện như sau:

Khoảng trọng lượng X, Y(kg)	< 35	[35;38)	[38;41)	[41;44)	[44;47)	[47;50)	[50;53)	≥ 53
Số trẻ em thành phố	0	2	8	13	20	15	12	8
Số trẻ em nông thôn	5	10	12	15	10	3	0	0

a) Với mức ý nghĩa $\alpha = 0,05$ có thể nói trọng lượng trung bình của trẻ em lứa tuổi lên 10 ở hai vùng có như nhau không hay vùng nào lớn hơn?

b) Ở lứa tuổi này ta xem trọng lượng ≥ 50kg là diện thừa cân, có nguy cơ béo phì, còn trọng lượng < 35kg là diện thiếu cân, có nguy cơ suy dinh dưỡng. Từ số liệu trên có thể kết luận:

– Tỷ lệ trẻ em lên 10 ở thành phố có nguy cơ béo phì cao hơn 25% hay không (với mức ý nghĩa $\alpha = 0,10$)?

– Tỷ lệ này thấp nhất là bao nhiêu % (với độ tin cậy 0,90)?

– Tỷ lệ trẻ em lên 10 ở nông thôn thuộc diện thiếu cân là bao nhiêu % ? Có thể nói tỷ lệ này nhỏ hơn 10% hay không (với $\alpha = 0,10$) ?

Giải:

a) Đây là so sánh hai trung bình EX, EY. Theo đề bài ta lần lượt giải hai bài toán sau:

H: $EX = EY$ / K: $EX \neq EY$; $\alpha = 0,05$.

Nếu chấp nhận $EX = EY$ thì bài toán dừng lại. Nếu chấp nhận $EX \neq EY$ thì ta cần làm tiếp bài toán thứ hai với đối thiết chọn $EX > EY$ hoặc $EX < EY$. Ta căn cứ vào độ lớn của hai ước lượng điểm $\overline{X}$ và $\overline{Y}$. Tính ra ta thấy $\overline{X} > \overline{Y}$ cho nên ta chọn đối thiết $EX > EY$.

Song cả hai bài toán cần phải làm nói trên, khi tính $\overline{X}$, $\overline{Y}$ ta thấy $\overline{X} > \overline{Y}$, do đó có thể chọn ngay bài toán

H: EX = EY/K : EX > EY; $\alpha = 0,05$.

Để giải các bài toán so sánh hai trung bình này ta dùng trường hợp c) vì DX, DY chưa biết, không có giả thiết chuẩn, nhưng $n_1 = 78$, $n_2 = 55$, có thể tạm chấp nhận là đủ lớn.

Tính toán ta được:

$n_1 = 78$; $\overline{X} = 46,577$; $s_x = 4,6569$; $\hat{s}_x = 4,6870$.

$n_2 = 55$; $\overline{Y} = 40,809$; $s_y = 4,076$; $\hat{s}_y = 4,1135$.

H: EX = EY / K : EX > EY; $\alpha = 0,05$; $u(0,05) = 1,65$

Miền tiêu chuẩn thích hợp là miền S_2(II.20).

$$u = \frac{46,577 - 40,809}{\sqrt{\frac{4,687^2}{78} + \frac{4,1135^2}{55}}} \approx 7,51 > 1,65.$$

Miền S_2(II.20) xảy ra → Ta bác bỏ EX = EY và chấp nhận EX > EY. Vậy trọng lượng trẻ em lên 10 ở thành phố lớn hơn ở nông thôn.

b) Gọi p_1 là tỷ lệ trẻ em béo phì ở thành phố. Ta có:

$$p_1^* = \frac{20}{78} = 0,2564$$

H: $p_1 = 0,25$ / K : $p_1 > 0,25$; $\alpha = 0,10$; $u(0,10) = 1,28$

Ta có: $$u = \frac{0,2564 - 0,25}{\sqrt{0,25 \cdot 0,75}} \cdot \sqrt{78} = 0,151 < 1,28$$

Miền $\overline{S}_2$ (II.11) xảy ra nên ta không chấp nhận p_1 cao hơn 25%.

Với độ tin cậy 90% có thể nói tỷ lệ trẻ béo phì p_1 thấp nhất là:

$$0,2564 - 1,65.\frac{\sqrt{0,2564\ .\ 0,7436}}{\sqrt{78}} = 0,1861 = 18,61\%.$$

Gọi p_2 là tỷ lệ trẻ lên 10 ở nông thôn thuộc diện thiếu cân. Ta có:

$$p_2^* = \frac{5}{55} = 0,0909 = 0,091.$$

Ta ước lượng cho p_2 là 9,1%.

H: $p_2 = 0,10$ / K : $p_2 < 0,10$; $\alpha = 0,10$; $u(0,10) = 1,28$

Ta có: $u = \frac{0,091 - 0,10}{\sqrt{0,1\ .\ 0,9}}\sqrt{55} = -0,222 > -1,28.$

Miền $\bar{S}_3$ (II.12) xảy ra nên ta không chấp nhận tỷ lệ p_2 nhỏ hơn 10% mặc dù ước lượng điểm cho nó là 9,1% nhỏ hơn 10%.

Trở lại bài toán so sánh hai giá trị trung bình. Ba trường hợp giải đã trình bày trên đòi hỏi các giả thiết về tính chuẩn, về phương sai (đã biết hoặc bằng nhau) và cỡ mẫu lớn, do đó khó dùng. Vì vậy, người ta đã xây dựng lớp các tiêu chuẩn phi tham số, mà đối với lớp này các giả thiết trên không đòi hỏi, do đó dễ dùng. Dưới đây giới thiệu hai tiêu chuẩn đại diện trong lớp tiêu chuẩn này.

d) Tiêu chuẩn phi tham số Mann - Whitney

Định nghĩa hạng (Rank):

Giả sử ta có dãy số liệu $x_1, x_2, \ldots, x_n$. Ta sắp xếp chúng theo thứ tự tăng dần: $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$.

– Nếu trong dãy, phần tử $x_{(i)}$ đứng một mình (không có giá trị nào trùng với nó) thì Rank $(x_{(i)}) = i$.

– Nếu có hai giá trị trùng nhau: $x_{(i-1)} < x_{(i)} = x_{(i+1)} < x_{(i+2)}$ thì Rank $(x_{(i)})$ = Rank $(x_{(i+1)}) = \frac{1}{2}(i + i + 1)$.

– Nếu có ba giá trị trùng nhau:

$$x_{(i-1)} < x_{(i)} = x_{(i+1)} = x_{(i+2)} < x_{(i+3)}$$

thì:

$$\text{Rank } (x_{(i)}) = \text{Rank } (x_{(i+1)}) = \text{Rank } (x_{(i+2)}) = \frac{1}{3}(i + i + 1 + i + 2).$$

– Nếu có 4, 5,... các giá trị trùng nhau thì cách xác định hạng được làm tương tự.

Giả sử ta có hai mẫu: $x_1, x_2, \ldots, x_{n_1}$ và $y_1, y_2, \ldots, y_{n_1}$ được rút ra một cách độc lập với nhau từ hai tập hoặc từ hai biến nào đó.

H : EX = EY / K : EX ≠ EY ; α

(Trong tiêu chuẩn phi tham số giả thiết H thường được phát biểu là: Hai mẫu thuần nhất, do đó đối thiết K là: Hai mẫu không thuần nhất. Ở đây khái niệm thuần nhất chỉ tương đương với sự bằng nhau của hai giá trị trung bình).

Các bước giải bài toán này như sau:

– Gộp hai mẫu thành một mẫu chung với số phần tử $n = n_1 + n_2$.

– Tính hạng của mỗi phần tử trong mẫu gộp.

– Tính tổng hạng của các phần tử thuộc từng mẫu:

$$R_1 = \sum_i \text{Rank}(x_{(i)});\ R_2 = \sum_j \text{Rank}(y_{(j)})$$

– Tính: $U_1 = n_1 n_2 + \frac{n_1(n_1 - 1)}{2} - R_1$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 - 1)}{2} - R_2$$

– Gọi U là U_1 hoặc U_2. Tính:

$$EU = \frac{n_1 n_2}{2};\ DU = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

Miền tiêu chuẩn của bài toán:

$$S = \left\{ \frac{|U - EU|}{\sqrt{DU}} \geq u\left(\frac{\alpha}{2}\right) \right\} \qquad \text{(II.22)}$$

Điểm hay nhất trong tiêu chuẩn này là các tác giả đã tìm được phân phối của U, đó là phân phối gần chuẩn với tham số EU và DU như trên. Từ kết quả này dẫn đến miền S sẽ là miền xảy ra với xác suất nhỏ α. (Trở lại nhận xét ở II.4.2).

Ví dụ II.14: Thời gian tự học trong một tuần của 12 sinh viên khoa A và 15 sinh viên khoa B được thống kê lại như sau (đơn vị: giờ):

Khoa A: 18 15 24 23 30 12 15 24 35 30 18 20

Khoa B: 19 18 24 25 30 36 28 25 30 12 14 28 22 28 20

Với mức ý nghĩa α = 0,02 thử xem thời gian tự học của sinh viên hai khoa thực chất có như nhau không?

Giải:

Đây là bài toán so sánh hai thời gian tự học trung bình. Trường hợp a), b), c) đều không thích hợp. Ta chỉ có thể dùng tiêu chuẩn Mann - Whitney. Gọi T_A, T_B là thời gian tự học của sinh viên khoa A và khoa B.

H: $ET_A = ET_B$ / K: $ET_A \neq ET_B$; $\alpha = 0,02$

Thuật toán trình bày trên là dùng cho máy tính. Còn khi tính toán bằng tay, ta thay đổi một chút: Sắp xếp từng mẫu riêng theo thứ tự, nhưng khi tính hạng thì tính hạng trong mẫu gộp, sau đó tính tổng R_1, R_2 sẽ dễ dàng hơn nhiều.

T_A	Rank $(t^A_{(i)})$
12	1,5
15	4,5
15	4,5
18	7
18	7
20	10,5
23	13
24	15
24	15
30	23,5
30	23,5
35	26
R_1 =	151,0

T_B	Rank $(t^B_{(j)})$
12	1,5
14	3
18	7
19	9
20	10,5
22	12
24	15
25	17,5
25	17,5
28	20
28	20
28	20
30	23,5
30	23,5
36	27
R_2 =	227,0

$n_1 = 12; n_2 = 15$

$$U = U_1 = 12.15 + \frac{12.13}{2} - 151 = 107$$

$$EU = \frac{12.15}{2} = 90; \quad DU = \frac{12.15.28}{12} = 420 \approx (20,49)^2$$

$$\frac{|107 - 90|}{20,49} = 0,83; \quad u(0,01) = 2,33$$

$0,83 < 2,33 \rightarrow$ miền $\bar{S}$ (II.22) xảy ra $\rightarrow$ Ta chấp nhận $ET_A = ET_B$, tức là thời gian tự học của sinh viên hai khoa thực chất là như nhau.

e) Tiêu chuẩn Wilcoxon

Ta vẫn xét bài toán kiểm định giả thiết H và đối thiết K như trường hợp Mann - Whitney, nhưng đối với hai mẫu phụ thuộc (hai mẫu cho theo từng cặp tương ứng với nhau).

Giả sử ta có hai mẫu $x_1, x_2, \ldots, x_n$

$y_1, y_2, \ldots, y_n$

H : $EX = EY$ / K : $EX \neq EY$; α

Các bước làm như sau:

– Tính hiệu $d_i = x_i - y_i$; $i = \overline{1,n}$

Loại bỏ các giá trị $d_i = 0$, gọi n^+ là các số $d_i \neq 0$.

– Tính Rank $\{|d_i|\}$ trong số n^+ giá trị khác 0.

$$\text{Tính } T^+ = \sum_{i:d_i>0} \text{Rank}\{|d_i|\}; \quad T^- = \sum_{i:d_i<0} \text{Rank}\{|d_i|\}$$

Gọi T là T^+ hoặc T^-.

– Tính: $ET = \frac{n^+(n^+ + 1)}{4}$; $DT = \frac{n^+(n^+ + 1)(2n^+ + 1)}{24}$

Miền tiêu chuẩn cần tìm:

$$S = \left\{ \frac{|T - ET|}{\sqrt{DT}} \geq u\left(\frac{\alpha}{2}\right) \right\} \qquad (II.23)$$

Giải thích về ý chứng minh: Tương tự như thống kê U, với $n^+ \geq 8$ tác giả chứng minh được rằng T có phân phối chuẩn, suy ra miền S(II.23) chỉ xảy ra với xác suất nhỏ α. (Trở lại nhận xét ở II.4.2 trang 106 – 107).

Ví dụ II.15: Gọi X, Y là năng suất theo giờ của công nhân trước và sau khi được nâng lương. Một mẫu ngẫu nhiên từ 20 công nhân cho ta kết quả sau:

Công nhân	A	B	C	D	E	F	G	H	I	J
X	90	83	70	64	85	86	91	66	72	60
Y	88	87	73	69	84	86	94	68	76	55
Công nhân	K	L	M	N	O	P	Q	R	S	T
X	75	84	71	80	70	85	65	75	75	65
Y	76	86	72	90	75	83	75	82	70	67

Với mức ý nghĩa $\alpha = 0,05$, hỏi việc nâng lương có tác dụng làm tăng năng suất của công nhân hay không?

Giải:

Ta nhận thấy số liệu được cho theo từng cặp tương ứng; (So sánh giá trị x và y tương ứng với một công nhân sẽ cho ta thông tin tăng, giảm năng suất của công nhân này. Nhưng lấy giá trị X của B so với giá trị Y của D thì vô nghĩa, chẳng cho thông tin gì). Vì vậy, trường hợp thích hợp duy nhất là tiêu chuẩn Wilcoxon.

H: $EX = EY$ / K: $EX \neq EY$; $\alpha = 0,05$; $u(0,05) = 1,65$

d_i	Rank $\{\|d_i\|\}$	d_i	Rank $\{\|d_i\|\}$
2	6	– 1	2
– 4	11,5	– 2	6
– 3	9,5	– 1	2
– 5	14,5	– 10	18,5
1	2	– 5	14,5
0		2	6
– 3	9,5	– 10	18,5
– 2	6	– 7	17
– 4	11,5	5	14,5
5	14,5	– 2	6

Để tính Rank $\{|d_i|\}$ ta làm như sau:

$|d_i| = 1$ có 3 giá trị, xếp thứ 1, 2, 3 → Rank $\{|d_i|\} = 2$.

$|d_i| = 2$ có 5 giá trị, xếp thứ 4, 5, 6, 7, 8 → Rank $\{|d_i|\} = 6$.

$|d_i| = 3$ có 2 giá trị, xếp thứ 9, 10.

$|d_i| = 4$ có 2 giá trị, xếp thứ 11, 12.

$|d_i| = 5$ có 4 giá trị, xếp thứ 13, 14, 15, 16.

$|d_i| = 7$ có 1 giá trị, xếp thứ 17.

$|d_i| = 10$ có 2 giá trị, xếp thứ 18, 19.

$n^+ = 19$.

$T = T^+ = 6 + 2 + 14,5 + 6 + 14,5 = 43$.

$$ET = \frac{19.20}{4} = 95;\ DT = \frac{19.20.39}{24} = 617,5 \approx (24,85)^2$$

$$\rightarrow \frac{|43-95|}{24,85} = 2,09 > 1,65 = u(0,05)$$

Miền S(II.23) xảy ra → Ta bác bỏ EX = EY và chấp nhận EX ≠ EY.

Nhận xét: Trong hai tiêu chuẩn phi tham số Mann – Whitney và Wilcoxon, ta chỉ có một đối thiết EX ≠ EY. Do đó, nếu đã chấp nhận EX ≠ EY, để so sánh EX > EY hoặc EX < EY ta căn cứ vào hai ước lượng điểm $\bar{X}$ và $\bar{Y}$. Nếu $\bar{X} > \bar{Y}$ thì kết luận EX > EY, nếu $\bar{X} < \bar{Y}$ thì kết luận EX < EY.

Áp dụng vào bài toán này, rõ ràng $\bar{X} < \bar{Y}$ (vì trong 19 giá trị thì có 13 giá trị $x_i < y_i$ ($d_i < 0$), chỉ có 6 giá trị $x_i > y_i$ ($d_i > 0$), cho nên ta kết luận EX < EY, nghĩa là về trung bình thì năng suất của công nhân có tăng lên sau khi tăng lương.

II.4.5. So sánh hai tỷ lệ (hai xác suất)

Gọi p là P(X∈A) hoặc p là tỷ lệ nào đó. p_1, p_2 là giá trị của p tương ứng ở hai tập hoặc hai biến nào đó (tất nhiên p_1, p_2 chưa biết).

$$H: p_1 = p_2 \,/\, K \begin{cases} p_1 \neq p_2 \\ p_1 > p_2 \\ p_1 < p_2 \end{cases} ;\ \alpha$$

Giả sử ta có hai mẫu với cỡ mẫu n_1, n_2 đủ lớn. Từ đó ta xác định được m_1, m_2 là số lần xảy ra biến cố A trong hai mẫu đã cho.

Tính:

$$u = \frac{\frac{m_1}{n_1} - \frac{m_2}{n_2}}{\sqrt{\frac{m_1+m_2}{n_1+n_2}\left(1-\frac{m_1+m_2}{n_1+n_2}\right).\frac{n_1+n_2}{n_1.n_2}}} = \frac{p_1^* - p_2^*}{\sqrt{p^*(1-p^*).\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

trong đó: p_1^*, p_2^* là hai ước lượng điểm quen thuộc, còn $p^* = \dfrac{m_1 + m_2}{n_1 + n_2}$.

Hai biểu thức trên tương đương nhau, nên bạn đọc có thể tuỳ ý chọn dùng một trong hai biểu thức đó.

Các miền tiêu chuẩn tương ứng là:

$$S_1 = \left\{ |u| \geq u\left(\frac{\alpha}{2}\right) \right\} \qquad \text{(II.24)}$$

$$S_2 = \{ u \geq u(\alpha) \} \qquad \text{(II.25)}$$

$$S_3 = \{ u \leq -u(\alpha) \} \qquad \text{(II.26)}$$

Ví dụ II.16: Để thăm dò ý kiến của nhân dân về một điều khoản nào đó, người ta chọn ra hai mẫu đại diện ở thành thị và ở nông thôn.

Ở thành thị: $n_1 = 500$; $m_1 = 320$ ý kiến ủng hộ.

Ở nông thôn: $n_2 = 400$; $m_2 = 300$ ý kiến ủng hộ.

Với mức ý nghĩa $\alpha = 0{,}05$ có thể kết luận người dân ở nông thôn ủng hộ điều khoản này cao hơn ở thành thị hay không?

Giải:

Ở đây không thể lấy số người ủng hộ để so sánh được vì nó không thuộc mô hình của bài toán nào mà ta đã xét. Để trả lời câu hỏi đặt ra ta dùng tỷ lệ người ủng hộ. Gọi p_t và p_n là tỷ lệ người ủng hộ ở thành thị và ở nông thôn tương ứng.

H: $p_t = p_n$ / K: $p_t < p_n$; $\alpha = 0{,}05$

Ta có: $u = \dfrac{\dfrac{320}{500} - \dfrac{300}{400}}{\sqrt{\dfrac{620}{900}\left(1 - \dfrac{620}{900}\right).\dfrac{900}{500.400}}} = -3{,}55.$

– 3,55 < –1,65 → Miền S_3(II.26) xảy ra → ta bác bỏ $p_t = p_n$ và chấp nhận $p_t < p_n$, nghĩa là có thể nói người dân ở nông thôn ủng hộ điều khoản này cao hơn người dân ở thành thị.

II.4.6. Tiêu chuẩn phù hợp χ^2

Bài toán đặt ra như sau:

Ta có k tỷ lệ cho trước $p_1, p_2, ..., p_k$;

$\sum_{i=1}^{k} p_i = 100\% = 1$ (Khác với các tình huống đã xét trước đây, các giá trị p_i đã biết, không phải tìm nữa).

Mặt khác, chúng ta có một mẫu cỡ n. Hỏi rằng số liệu mẫu có phù hợp với k tỷ lệ đã cho hay không ?

Giả thiết H: Số liệu mẫu phù hợp với k tỷ lệ đã cho.

Đối thiết K: Số liệu mẫu không phù hợp với k tỷ lệ đã cho.

Để giải quyết bài toán, các bước làm như sau:

– Từ số liệu mẫu ta xác định các tần số $m_1, m_2, \ldots, m_k$; $\sum_{i=1}^{k} m_i = n$, trong đó m_i là số lần xảy ra tình huống tương ứng với tỷ lệ p_i.

– Tính tổng: $\chi^2 = \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i} = \frac{1}{n}\sum_{i=1}^{n} \frac{m_i^2}{p_i} - n$

Miền bác bỏ H: $S = \{ \chi^2 \geq \chi^2_{k-1}(\alpha) \}$ (II.27)

Giải thích về ý chứng minh: Nếu giả thiết H đúng thì Pearson đã chứng minh được rằng tổng χ^2 có phân phối dần đến phân phối χ^2_{k-1}, coi như χ^2 có phân phối χ^2_{k-1}, do đó miền S (II.27) sẽ xảy ra chỉ với xác suất $\leq \alpha$ (α nhỏ), vì vậy thấy S xảy ra, ta nghi ngờ tính đúng đắn của H.

Nhận xét: Tiêu chuẩn này được dùng tốt nếu các tần số m_i không quá bé ($m_i \geq 5$; $\forall i$). Do đó nếu trong các tần số m_i, có giá trị nào < 5 thì ta gộp tần số này với tần số lân cận (do đó giá trị p_i cũng được gộp lại tương ứng) để thỏa mãn $m_i \geq 5$; $\forall i$. Sau bước này sẽ cho ta giá trị k là bao nhiêu, từ đó ta biết được tham số (số bậc tự do) (k – 1) của phân phối giới hạn.

Ví dụ II.17: Theo tổng kết thì trước khi có Nghị định 36/CP tỷ lệ các vụ tai nạn giao thông đường bộ do người đi bộ, do xe đạp, do mô tô (xe máy), do ô tô gây ra ở thành phố Z tương ứng là 10%, 15%, 60% và 15%. Sau 3 tháng thực hiện Nghị định 36/CP, ở thành phố Z đã xảy ra 250 vụ tai nạn giao thông đường bộ, trong đó có 40 vụ do lỗi của người đi bộ, 60 vụ do lỗi của người đi xe đạp, 120 vụ do xe máy gây ra và 30 vụ do ô tô gây ra.

a) Với mức ý nghĩa $\alpha = 0,05$ có thể nói rằng sau Nghị định 36/CP nguyên nhân gây ra tai nạn giao thông đường bộ đã thay đổi so với trước hay không? Rút ra ý nghĩa thực tiễn gì từ kết luận nhận được?

b) Với mức ý nghĩa $\alpha = 0,05$ có thể nói tỷ lệ các vụ tai nạn do người đi bộ và người đi xe đạp tăng lên, còn do xe máy và ô tô là giảm đi hay không?

Giải:

a) Ta có 4 tỷ lệ đã cho và một mẫu với $n = 250$, $m_1 = 40$, $m_2 = 60$, $m_3 = 120$, $m_4 = 30$. Các m_i đều ≥ 30 (vượt xa yêu cầu > 5); $k = 4$.

H: Số liệu mẫu phù hợp với các tỷ lệ đã cho, hoặc tương dương: Nguyên nhân gây ra tai nạn giao thông đường bộ không thay đổi so với trước.

K: Số liệu mẫu không phù hợp với các tỷ lệ đã cho, hoặc tương dương: Nguyên nhân gây ra tai nạn đã thay đổi so với trước.

Tính tổng:

$$\chi^2 = \frac{(40-250.0,10)^2}{250.0,10} + \frac{(60-250.0,15)^2}{37,5} + \frac{(120-250.0,6)^2}{150} +$$

$$+ \frac{(30-250.0,15)^2}{37,5} = 39,0$$

Tra bảng ta được $\chi^2_3(0,05) = 7,81$.

$39,0 > 7,81 \rightarrow$ Ta bác bỏ H, chấp nhận K, nghĩa là số liệu mới về các vụ tai nạn không phù hợp với 4 tỷ lệ đã có trước đây, hay nguyên nhân gây ra tai nạn giao thông đường bộ đã thay đổi so với trước khi có Nghị định 36/CP.

Từ kết luận trên ta thấy sự chuyển biến về ý thức chấp hành luật giao thông ở 4 đối tượng tham gia giao thông đường bộ là khác nhau. Đối tượng chuyển biến nhiều thì tỷ lệ số vụ tai nạn do họ gây ra giảm đi, còn đối tượng ít chuyển biến thì tỷ lệ số vụ tai nạn do họ gây ra sẽ tăng lên so với tương quan chung. (Có bạn sẽ suy ra rằng: số vụ đã thay đổi, số vụ tăng lên, Nghị định 36/CP chưa có tác dụng,... Nhưng theo đầu bài thì không có cơ sở nào nghĩ và suy luận như thế được.

b) Gọi $p_{đb}$, $p_{xđ}$, p_{xm}, p_{ot} là 4 tỷ lệ tai nạn sau Nghị định 36/CP do người đi bộ, xe đạp, xe máy và ô tô gây ra. 4 tỷ lệ thực này chưa biết, ta cần so sánh chúng với 4 tỷ lệ đã cho. Ta có 4 bài toán kiểm định về tỷ lệ với các miền tiêu chuẩn (II.11), (II.12):

H_1: $p_{đb} = 0,10$ / K_1: $p_{đb} > 0,10$; $\alpha = 0,05$; $u(0,05) = 1,65$.

$$u = \frac{\frac{40}{250} - 0,1}{\sqrt{0,1.0,9}} . \sqrt{250} = 3,162 > 1,65 \rightarrow \text{Chấp nhận } p_{đb} > 10\%.$$

H_2: $p_{xđ} = 0,15$ / K_2: $p_{xđ} > 0,15$; $\alpha = 0,05$

$$u = \frac{\frac{60}{250} - 0,15}{\sqrt{0,15.0,85}} . \sqrt{250} = 3,953 > 1,65 \rightarrow$$ Chấp nhận $p_{xđ} > 15\%$.

H_3: $p_{xm} = 0,60$ / K_3 : $p_{xm} < 0,60$; $\alpha = 0,05$

$$u = \frac{\frac{120}{250} - 0,6}{\sqrt{0,6.0,4}} . \sqrt{250} = -3,872 < -1,65 \rightarrow$$ Chấp nhận $p_{xm} < 60\%$.

H_4: $p_{ot} = 0,15$ / K_4 : $p_{ot} < 0,15$; $\alpha = 0,05$

$$u = \frac{\frac{30}{250} - 0,15}{\sqrt{0,15.0,85}} . \sqrt{250} = -1,32 > -1,65 \rightarrow$$ Chấp nhận $p_{xm} = 15\%$.

Chưa có cơ sở để kết luận $p_{ot} < 15\%$.

Qua các kết luận của câu b) chúng ta càng thấy rõ là nguyên nhân gây tai nạn đã thay đổi. Xe máy và ô tô sợ bị phạt nặng nên có ý thức hơn, nhưng người đi bộ và xe đạp chưa bị phạt nên ít chuyển biến. Ta cần tuyên truyền, nhắc nhở hai đối tượng này. (Các số liệu mẫu ở trên chỉ là giả định, cho nên các kết luận rút ra cũng chỉ mang tính giả định).

II.4.7. Kiểm tra tính độc lập

Bài toán đặt ra là: Chúng ta có hai đại lượng (hai đặc trưng xã hội nào đó) X và Y. Thử xem 2 đại lượng này độc lập với nhau hay có sự phụ thuộc lẫn nhau.

H: X và Y độc lập.

K: X và Y phụ thuộc.

Giả sử ta có mẫu ngẫu nhiên cỡ n về 2 biến ngẫu nhiên X và Y: (x_i, y_i) ; $i = \overline{1,n}$

(Mỗi quan sát cho hai giá trị x_i, y_i tương ứng với nhau). Giá trị của X và Y có thể là định lượng, cũng có thể là định tính (chẳng hạn phân loại chất lượng, các màu sắc, phân chia các cá thể,...). Giả sử biến X nhận r giá trị (hoặc r khoảng), biến Y nhận s giá trị (hoặc s khoảng). Khi đó, để giải quyết bài toán đặt ra, các bước làm như sau:

– Từ mẫu đã cho ta xác định số lần biến X nhận giá trị $x_{(i)}$, biến Y nhận giá trị $y_{(j)}$. Ký hiệu số này là n_{ij}, $i = \overline{1,r}$, j = 1, s. Ta có một bảng gồm r.s ô:

Y / X	$y_{(1)}$	. . .	$y_{(j)}$	. . .	$y_{(s)}$	Tổng hàng: hg1
$x_{(1)}$	n_{11}		n_{ij}		n_{1s}	hg1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$x_{(i)}$	n_{i1}		n_{ij}		n_{is}	hgi
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$x_{(n)}$	n_{r1}		n_{rj}		n_{rs}	hgr
Tổng cột: cotj	cot1		cotj		cots	n

– Tính tổng theo hàng, tổng theo cột:

$$hgi = \sum_{j=1}^{s} n_{ij} \; ; \quad cot\, j = \sum_{i=1}^{r} n_{ij}$$

(Để bạn đọc làm quen với việc lấy tổng theo hai chỉ số, ta có:

Tổng theo mọi ô:

$(n_{11} + ... + n_{1s}) + (n_{21} + ... + n_{2s}) + ... + (n_{r1} + ... + n_{rs}) = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = n$

Hệ thức trên tương đương với:

$$hg1 + hg2 + ... + hgr = \sum_{i=1}^{r} hgi = n$$

$$\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = \sum_{i=1}^{r}\left(\sum_{j=1}^{s} n_{ij}\right) = \sum_{i=1}^{r} hgi = n$$

$$\sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij} = \sum_{j=1}^{s}\left(\sum_{i=1}^{r} n_{ij}\right) = \sum_{j=1}^{s} Cotj = n\ )$$

– Mỗi ô (i, j) ta tính một con số bằng: $\frac{hgi.cot j}{n}$. Đặt con số này ở trong ngoặc ( ) để phân biệt với con số n_{ij} đã có trước đó.

– Tính tổng:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{hgi.cot j}{n}\right)^2}{\frac{hgi.cot j}{n}} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{hgi.cot j} - 1\right)$$

Miền bác bỏ tính độc lập của X và Y là:

$$S = \left\{ \chi^2 \geq \chi^2_{(r-1)(s-1)}(\alpha) \right\} \qquad (II.28)$$

Lời giải vừa nêu trên là trường hợp riêng khi đưa về áp dụng tiêu chuẩn phù hợp χ^2. Nếu H đúng thì tổng χ^2 có phân phối khi bình phương với tham số (r – 1)(s – 1). Do đó miền S (II.28) sẽ có xác suất nhỏ α. Vì thế, nếu S(II.28) xảy ra ta sẽ bác bỏ H. Thực hiện như vậy sẽ thỏa mãn yêu cầu 1' và 2' của bài toán kiểm định giả thiết.

Nếu dùng tổng χ^2 dạng sau (nhận được do khai triển bình phương và rút gọn từ tổng đầu) thì ta bỏ qua bước 3.

Ví dụ II.18: Nghiên cứu về sự phụ thuộc giữa đạo đức của trẻ ở tuổi vị thành niên và hoàn cảnh gia đình của họ, ta có một mẫu được điều tra ngẫu nhiên như sau:

Hoàn cảnh gia đình / Phân loại đạo đức	Bố hoặc mẹ đã mất	Bố mẹ ly hôn	Còn cả bố mẹ	Tổng hàng
Ngoan	20	15	50	85
Hư	8	12	10	30
Phạm tội	6	9	5	20
Tổng cột	34	36	65	135

Với mức ý nghĩa $\alpha = 0,10$ có thể kết luận tình trạng đạo đức của trẻ ở tuổi vị thành niên phụ thuộc vào hoàn cảnh gia đình của họ hay không?

Giải:

H: Tình trạng đạo đức và hoàn cảnh gia đình là độc lập.

K: Tình trạng đạo đức và hoàn cảnh gia đình là phụ thuộc.

Để tính toán, ta tính tổng hàng và tổng cột. Để đỡ phải lập lại bảng số liệu đã cho ta dùng luôn bảng trên.

$$\chi^2 = 135\left(\frac{20^2}{34.85}+\frac{15^2}{36.85}+\frac{50^2}{65.85}+\frac{8^2}{34.30}+\frac{12^2}{36.30}+\frac{10^2}{65.30}\right)$$

$$+135\left(\frac{6^2}{34.20}+\frac{9^2}{36.20}+\frac{5^2}{65.20}-1\right)$$

$= 135.0,0965 = 13,028.$

Tra bảng ta được $\chi^2_1(0,10) = 7,78.$

13,028 > 7,78 → Ta bác bỏ tính độc lập và chấp nhận tình trạng đạo đức của trẻ vị thành niên phụ thuộc vào hoàn cảnh gia đình của họ.

II.4.8. So sánh nhiều tỷ lệ

Trong II.4.5. ta đã so sánh hai tỷ lệ. Bây giờ ta xét bài toán so sánh s tỷ lệ ($s \geq 2$).

Gọi $p = P(A)$ hoặc là một tỷ lệ nào đó đang cần nghiên cứu.

$p_1, p_2, ..., p_s$ là các giá trị chưa biết của p tương ứng ở s tập hoặc s biến.

H: $p_1 = p_2 = \ldots = p_s \qquad (s \geq 2)$

K: Các tỷ lệ không như nhau.

Để giải bài toán này ta đưa về áp dụng bài toán kiểm tra độc lập. Bởi lẽ khi xét một tỷ lệ chưa biết, chúng ta cần hai con số n và m,tức là tập số liệu mẫu sẽ chia làm hai lớp: số lần xảy ra biến cố A (m) và số lần không xảy ra A, tức là xảy ra $\bar{A}$ (n – m). Nếu hai dấu hiệu A, $\bar{A}$ độc lập với các đối tượng (các tập) thì tỷ lệ p đang xét sẽ như nhau đối với s đối tượng, còn nếu dấu hiệu A phụ thuộc vào các đối tượng thì tỷ lệ p đang xét sẽ thay đổi theo đối tượng. Nghĩa là, giả thiết H ở trên tương đương với tính độc lập của dấu hiệu A với s đối tượng, đối thiết K tương đương với tính phụ thuộc của A với s đối tượng. Vì vậy, lời giải của bài toán này hoàn toàn tương tự như bài toán kiểm tra độc lập (xin lưu ý là, ở đây r = 2 vì chỉ có hai dấu hiệu A và $\bar{A}$).

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{hgi.cot\,j}{n}\right)^2}{\frac{hgi.cot\,j}{n}} = n\left(\sum_{i=1}^{2}\sum_{j=1}^{s}\frac{n_{ij}^2}{hgi.cot\,j} - 1\right)$$

Miền bác bỏ giả thiết H sẽ là:

$$S = \left\{ \chi^2 \geq \chi^2_{s-1}(\alpha) \right\} \qquad (II.29)$$

Ví dụ II.19: Một hãng sản xuất ô tô muốn tìm hiểu xem có sự phụ thuộc nào giữa giới tính của người sở hữu ô tô với kiểu dáng của ô tô hay không. Một mẫu ngẫu nhiên gồm 2000 chủ sở hữu ô tô đã được chọn và phân loại như sau (xem [2]):

Kiểu dáng ô tô / Giới tính	I	II	III
Nam	350	270	380
Nữ	340	400	260

Với mức ý nghĩa $\alpha = 0{,}025$, tỷ lệ nữ dùng ba loại ô tô trên có như nhau không?

Giải:

Rõ ràng bài toán này có thể được giải theo bài toán kiểm tra tính độc lập cũng như theo mô hình so sánh nhiều tỷ lệ. Theo đề bài thì ta dùng bài toán so sánh ba tỷ lệ.

Gọi p_i là tỷ lệ nữ dùng ô tô kiểu dáng loại i ; i = 1, 2, 3.

H: $p_1 = p_2 = p_3$.

K: 3 tỷ lệ không như nhau; $\alpha = 0{,}025$.

Để tính tổng χ^2 ta dùng biểu thức đầu. Ta có bảng tính như sau:

Kiểu dáng ô tô / Giới tính	I	II	III	Tổng hàng
Nam	350 (345)	270 (335)	380 (320)	1000
Nữ	340 (345)	400 (335)	260 (320)	1000
Tổng cột	690	670	640	2000

$$\chi^2 = \frac{(350-345)^2 + (340-345)^2}{345} + \frac{(270-335)^2 + (400-335)^2}{335}$$
$$+ \frac{(380-320)^2 + (260-320)^2}{320}$$
$$= 0,145 + 25,224 + 22,5 = 47,869$$

$\chi^2_2(0,025) = 7,38 \rightarrow 47,869 > 7,38$

Vậy ta chấp nhận ba tỷ lệ khác nhau, hoặc kiểu dáng ô tô có phụ thuộc vào giới tính của chủ sở hữu.

II.5. TƯƠNG QUAN VÀ HỒI QUY ĐƠN

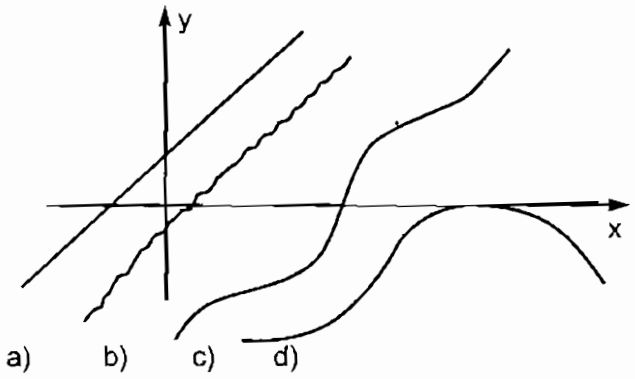
Đặt bài toán: Khi xét đồng thời hai biến ngẫu nhiên X và Y, như trong II.4.7 ta đã biết, hai biến này có thể độc lập với nhau hoặc hai biến này có sự phụ thuộc với nhau. Bây giờ ta tiếp tục xem xét tình huống hai biến có sự phụ thuộc với nhau. Hỏi rằng loại phụ thuộc là loại nào? Mức độ phụ thuộc là chặt hay lỏng? Biểu thức phụ thuộc như thế nào?

Có nhiều kiểu phụ thuộc, nhưng phổ biến nhất là dạng phụ thuộc hàm số: Y = f(X). Có rất nhiều dạng hàm f. Nhưng dạng đơn giản nhất là dạng bậc nhất Y = aX + b, hay là dạng tuyến tính, hay là dạng đường thẳng. Ở mức độ đơn giản của giáo trình chúng ta sẽ dừng lại nghiên cứu ở dạng đơn giản nhất: dạng tuyến tính (mặc dù cho đến nay trong Thống kê Toán học người ta đã nghiên cứu nhiều dạng phụ thuộc khác). Trong mục này chúng ta sẽ trả lời hai câu hỏi:

– Số đo mức độ phụ thuộc tuyến tính giữa hai biến là thế nào?

– Biểu thức phụ thuộc dạng tuyến tính giữa hai biến là gì?

Giả sử ta có bốn dạng phụ thuộc cụ thể được minh hoạ trong hình vẽ sau:



Hình II.7. Minh họa xu thế đường thẳng trong một số dạng phụ thuộc
a) Quan hệ giữa hai biến là quan hệ bậc nhất, 100% là đường thẳng.
b) Tính đường thẳng so với (a) là kém hơn,
nhưng so với (c) xu thế đường thẳng còn rõ hơn nhiều.
c) Xu thế đường thẳng so với (b) là kém hơn,
nhưng so với (d) tính đường thẳng của (c) lại rõ hơn.

Như vậy, rõ ràng có tồn tại một số đo, để đo mức độ đường thẳng trong sự phụ thuộc giữa hai biến. Số đo ấy chính là hệ số tương quan được trình bày dưới đây.

II.5.1. Hệ số tương quan

Định nghĩa: Ta xây dựng đại lượng ρ như sau:

$$\rho = \frac{E\{(X - EX)(Y - EY)\}}{\sqrt{DX}\ \sqrt{DY}} = \frac{E(XY) - EX.EY}{\sqrt{DX}\ \sqrt{DY}}$$

Đại lượng ρ được gọi là hệ số tương quan giữa hai biến ngẫu nhiên X và Y.

(Tất nhiên, để ρ xác định thì đòi hỏi $DX > 0$, $DY > 0$).

Tính chất: $-1 \le \rho \le 1$ hay $|\rho| \le 1$

$$\rho(aX + b, CY + d) = \rho(X, Y) \quad (\text{nếu } ac > 0)$$

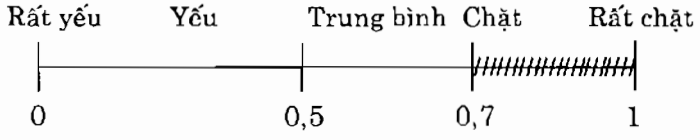
(Tính chất thứ hai bạn đọc có thể tự chứng minh được khi dùng các tính chất của kỳ vọng và phương sai $E(aX + b) = aEX + b$, $D(aX + b) = a^2DX$. Còn tính chất thứ nhất: $|\rho| \le 1$, thì chúng ta phải công nhận. Thống kê Toán học đã chứng minh điều đó).

Ý nghĩa: Hệ số tương quan ρ là một con số dùng để đo mức độ phụ thuộc tuyến tính giữa hai biến. Nếu $|\rho|$ càng lớn thì mức độ phụ thuộc tuyến tính giữa hai biến càng chặt. Đặc biệt, nếu $|\rho| = 1$ thì chắc chắn $Y = aX + b$, tức là $P(Y = aX + b) = 1$. Còn nếu $|\rho|$ càng nhỏ thì mức độ phụ thuộc tuyến tính giữa hai biến càng yếu. Nếu $|\rho| = 0$ thì hai biến không có sự phụ thuộc tuyến tính. Khi đó, người ta thường nói hai biến không tương quan.

Ta có: Nếu hai biến độc lập → hai biến không tương quan ($\rho = 0$).

Nhưng điều ngược lại không đúng, nghĩa là khi $\rho = 0$, hai biến không có sự phụ thuộc tuyến tính, song rất có thể chúng có sự phụ thuộc phi tuyến rất chặt, chứ không hẳn chúng độc lập với nhau.

Trong Thống kê Toán học, độ lớn của $|\rho|$ và mức độ phụ thuộc tuyến tính được định lượng như sau:



Trong trường hợp có sự phụ thuộc tuyến tính, nếu $\rho > 0$ thì hai biến là đồng biến, còn nếu $\rho < 0$ thì hai biến là nghịch biến.

Chúng ta công nhận những kết luận về ý nghĩa của ρ. Tất cả những điều này đều đã được các nhà Toán học chứng minh đầy đủ.

Hệ số tương quan mẫu

Trên thực tế chúng ta không biết được EX, DX, EY, DY,... Thông tin duy nhất chúng ta có là một mẫu ngẫu nhiên gồm n quan sát độc lập:

$$(x_i, y_i)\ ;\ i = \overline{1, n}$$

Nếu thực hiện việc thu gọn mẫu (Xem II.2.3f) ta được:

$$\begin{cases} (x_i, y_i) \\ m_i, \quad \sum_{i=1}^{k} m_i = n \end{cases}$$

Xuất phát từ mẫu đại diện, để nhận được ước lượng điểm cho ρ ta thay các số đặc trưng lý thuyết bởi các ước lượng điểm tương ứng của chúng, ta nhận được hệ số tương quan mẫu r như sau:

$$r = \frac{\overline{XY} - \bar{X}\bar{Y}}{s_x s_y} \qquad \text{(II.30)}$$

trong đó: $\overline{XY} = \frac{1}{n}\sum_{i=1}^{k} m_i x_i y_i$.

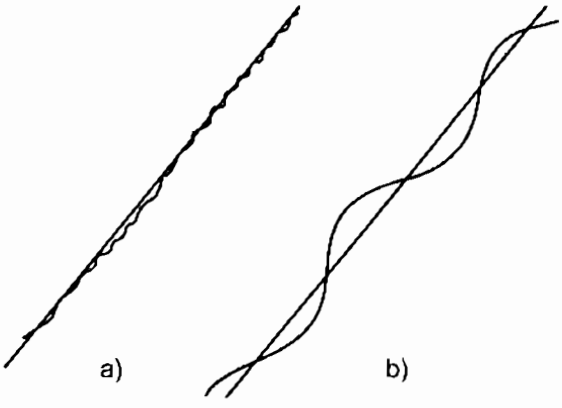
II.5.2. Đường hồi quy bình phương trung bình tuyến tính

Việc tìm biểu thức chính xác sự phụ thuộc Y = f(X) là công việc còn xa ở phía trước. Trong mục này chúng ta sẽ tìm một đường thẳng (dạng tuyến tính của Y theo X) để sao cho nó xấp xỉ tốt nhất đường cong Y = f(X) cần tìm.

Trong hình vẽ dưới đây: đường cong là đường Y = f(X) chưa biết, đang cần tìm; đường thẳng là đường xấp xỉ tốt

nhất đường Y = f(X). Rõ ràng đường thẳng như thế có tồn tại. Câu trả lời cho bài toán đặt ra là:

$$Y - EY = \rho\sqrt{\frac{DY}{DX}}\ (X - EX)$$



Hình II.8. Minh họa đường thẳng xấp xỉ tốt nhất đường cong.

Đó là đường thẳng xấp xỉ tốt nhất đường cong Y = f(X) cần tìm. Nó được gọi là đường hồi quy bình phương trung bình tuyến tính.

Tuy là đường xấp xỉ tốt nhất nhưng vẫn có sai số. Trên hình II.8: trường hợp (a) xấp xỉ tốt hơn, sai số nhỏ hơn trường hợp (b). Vậy, sai số bình phương trung bình mắc phải khi dùng đường hồi quy bình phương trung bình tuyến tính là:

$$\sigma^2_{Y/X} = DY\ (1 - \rho^2)$$

(Chúng ta bỏ qua việc chứng minh hai kết quả này mà chỉ cần hiểu rõ cách đặt bài toán và câu trả lời của nó. Thực ra, để chứng minh các kết quả này ta có thể dùng toán sơ cấp [xem 1]).

Nhìn vào công thức sai số $\sigma^2_{Y/X}$ ta thấy: sai số càng nhỏ, nghĩa là xấp xỉ càng tốt khi $|\rho|$ gần 1, tức là mức độ phụ thuộc tuyến tính rất chặt. Điều đó hoàn toàn phù hợp với những điều ta đã thấy từ hình II.8.

Hai công thức trên vẫn là công thức lý thuyết. Xuất phát từ mẫu đại diện, để nhận được ước lượng cho chúng, ta thay các số đặc trưng lý thuyết bởi ước lượng điểm tương ứng. Khi đó ta nhận được:

$$y - \bar{y} = r.\frac{s_y}{s_x}(x - \bar{x})$$

Tương tự: $x - \bar{x} = r.\frac{s_x}{s_y}(y - \bar{y})$ (II.31)

Đó là phương trình hai đường hồi quy bình phương trung bình tuyến tính thực nghiệm, gọi tắt là đường hồi quy tuyến tính thực nghiệm; còn ước lượng cho sai số bình phương trung bình là:

$$s^2_{y/x} = s^2_y(1 - r^2).$$

Tương tự: $s^2_{y/x} = s^2_x(1 - r^2)$. (II.32)

Nhìn vào các công thức (II.30), (II.31) và (II.32), xuất phát từ mẫu đại diện chúng ta chỉ cần hai lần tính $\bar{X}$ và s đối với hai mẫu (x_i, m_i) và (y_i, m_i); $i = \overline{1, k}$ (xem II.2.5); sau đó tính $\overline{XY} = \frac{1}{n}\sum_{i=1}^{k} m_i x_i y_i$. Lắp vào công thức, chúng ta nhận được các kết quả cần tìm.

Trong máy tính Casio fx 500 MS hoặc Casio 570M cũng có sẵn các chương trình để tính hệ số tương quan mẫu, xây dựng đường hồi quy

tuyến tính thực nghiệm và ước lượng sai số bình phương trung bình. Nếu có máy tính loại trên bạn đọc nên khai thác để tiện tính toán.

Có vài biểu thức khác nhau của r, $s^2_{y/x}$,... (tất nhiên kết quả là như nhau), nhưng các biểu thức dạng (II.30) – (II.32) là dễ nhớ và dễ làm.

Chỉ cần $s_x > 0$ (hoặc $s_y > 0$) là ta có thể lập được đường hồi quy tuyến tính thực nghiệm dạng (II.31). Nhưng lập được, không có nghĩa là dùng được. Vậy khi nào ta có thể dùng được đường hồi quy tuyến tính thực nghiệm (II.31) để tìm giá trị của biến này xấp xỉ theo giá trị của biến kia? Câu trả lời là khi $|r| \geq 0{,}70$; tức là khi mức độ phụ thuộc tuyến tính giữa hai biến là ở mức chặt trở lên.

Ví dụ II.20: Điều tra số trẻ em thất học (X) và số trẻ em hư (Y) ở 30 địa phương ta có các số liệu sau:

X \ Y	16	18	20	22	24	26	Tổng hàng
80	3	1					4
90		4	3	1			8
100		1	3	1			5
110			3	1	1		5
120				2	2	1	5
130					2	1	3
Tổng cột	3	6	9	5	5	2	30

a) Tính hệ số tương quan mẫu r.

b) Có nhận xét gì về sự phụ thuộc giữa số trẻ hư và số trẻ thất học qua kết quả r tính được?

c) Xây dựng đường hồi quy tuyến tính thực nghiệm của số trẻ hư theo số trẻ thất học.

d) Đường hồi quy nhận được có thể dùng làm gì? Vì sao? Cho ví dụ.

e) Ước lượng sai số bình phương trung bình khi dùng đường hồi quy trên.

Giải:

Số liệu được rút gọn ở dạng bảng hai lối vào (xem II.2.3.f)), ta có thể chuyển về dạng hàng hoặc dạng cột. Nhưng ở đây ta sẽ tính toán với số liệu được cho dưới dạng bảng hai lối vào. Tính tổng theo hàng và theo cột. (Để đỡ lập lại bảng số liệu mới ta dùng bảng đã cho). Các giá trị ở cột tổng hàng cho ta các tần số m_i ứng với các giá trị x_i của biến X:

X:	80	90	100	110	120	130
m_i:	4	8	5	5	5	3

Tương tự, các giá trị ở tổng cột cho ta các tần số m_i ứng với biến Y:

Y:	16	18	20	22	24	26
m_i:	3	6	9	5	5	2

Dùng máy tính Casio ta nhận được:

$\bar{X} = 102,667;$ $s_x = 15,691$

$\bar{Y} = 20,60;$ $s_y = 2,788;$ $\overline{XY} = 2152,667.$

(Nếu dùng Casio fx 500A, chỉ có chương trình tính $\bar{X}$ và s, để tính $\overline{XY}$ ta nên dùng chức năng M+ và MR (để xóa sạch số liệu cũ trong MR ta bấm SHIFT và DEL)).

a) Hệ số tương quan mẫu

$$r = \frac{2152,667 - 102,667 \times 20,6}{15,691 \times 2,788} = 0,8624.$$

b) Với r = 0,8624, từ ý nghĩa của hệ số tương quan ta thấy mức độ phụ thuộc tuyến tính giữa số trẻ hư và số trẻ

thất học là khá chặt. Hơn nữa sự phụ thuộc là đồng biến, nghĩa là số trẻ thất học càng nhiều thì số trẻ hư càng tăng.

c) Theo (II.31), đường hồi quy của số trẻ hư theo số trẻ thất học, hay đường hồi quy của biến y theo biến x là:

$$y - 20,6 = 0,8624.\ \frac{2,788}{15,691}\ (x - 102,667).$$

$$y = 0,1532x + 4,8714.$$

d) Đường hồi quy nhận được có thể dùng để dự đoán số trẻ hư khi cho biết số trẻ thất học, vì r = 0,8624 khá gần 1, sự phụ thuộc tuyến tính khá chặt cho nên xấp xỉ khá tốt, dự đoán sẽ khá tốt.

Ví dụ:

Nếu x = 90 thì $y_{\text{quan sát}} = \frac{1}{8}(18.4 + 20.3 + 22.1) = 19,25$; trong khi đó theo đường hồi quy thì $y_{\text{dự đoán}} = 0,1532.90 + 4,8714 = 18,6594$. Nghĩa là $y_{\text{quan sát}} = 19,25 \approx 19$, còn $y_{\text{dự đoán}} = 18,66 \approx 19$. Giá trị dự đoán khá sát với giá trị quan sát được.

Nếu x = 105 trẻ thất học, khi đó ta dự đoán sẽ có số trẻ hư là:

0,1532.105 + 4,8714 = 20,957, tức là gần 21 trẻ hư.

e) Sai số bình phương trung bình mắc phải khi dùng đường hồi quy trên để dự đoán giá trị của y là:

$$s^2_{y/x} = 2,788^2\ (1 - 0,8624^2) = 1,9922 = (1,41)^2.$$

BÀI TẬP CHƯƠNG II

1. Giả sử 100 sinh viên đăng ký vào lớp Thống kê và sau đây là kết quả thi của họ (X, tính theo thang điểm 100):

77	44	49	33	38	33	76	55	68	39
44	59	36	55	47	61	53	32	65	51
29	41	31	45	83	58	73	47	40	26
59	43	66	44	41	25	39	72	37	55
34	47	66	53	55	58	49	45	61	41
55	92	83	77	45	62	45	36	78	48
54	50	51	66	80	73	57	61	56	50
45	82	71	48	46	69	38	72	56	64
38	45	51	44	41	68	45	92	43	12
37	16	44	57	63	71	40	64	57	51

(Số liệu được trích từ [1])

a) Sắp xếp các số liệu trên theo thứ tự tăng dần.

b) Gộp các giá trị trùng nhau để có được mẫu thu gọn. Điểm nào có tần số (m_i) cao nhất?

c) Thu gọn số liệu trên bằng cách ghép khoảng (chia làm 10 khoảng bởi các điểm chia 10,5; 19,5; 28,5; ... 82,5; 91,5 và 100,5).

d) Xác định Med từ số liệu ở câu a) và từ số liệu ở câu c).

e) Xác định Q_1, Q_3.

f) Tính $\bar{X}$, s, ŝ từ số liệu ở câu b).

g) Với độ tin cậy 0,95, hãy chỉ ra ước lượng khoảng cho EX.

h) Gọi p = P{X ≥ 80} (tỷ lệ sinh viên đạt điểm giỏi).

Chỉ ra ước lượng điểm và ước lượng khoảng cho p với độ tin cậy là 0,90.

2. Công đoàn nhà trường muốn tìm hiểu xem mức sống của các gia đình ra sao, họ đã thống kê thu nhập hằng năm của các gia đình ở một khu tập thể. Kết quả được cho như sau:

Mức thu nhập X (triệu đồng)	Số gia đình
(6,0 – 7,0]	4
(7,0 – 8,0]	5
(8,0 – 9,0]	6
(9,0 – 10,0]	8
(10,0 – 11]	12
(11,0 – 12]	14
(12,0 – 13]	15
(13,0 – 14]	18
(14,0 – 15]	20
(15,0 – 16]	19
(16,0 – 17]	18
(17,0 – 18]	16
(18,0 – 19]	14
(19,0 – 20]	12
(20,0 – 21]	10
(21,0 – 22]	8
(22,0 – 23]	6
(23,0 – 24]	5
> 24	5
	215

Giả sử thu nhập của mọi gia đình trong khu tập thể này có thể đại diện chung cho các gia đình trong toàn trường; $X \simeq N(\mu, \sigma^2)$.

a) Xác định Med, Q_1, Q_3.

b) Xác định $\bar{X}$, $\hat{s}^2$, s, $\hat{s}$.

c) Ước lượng điểm của μ và ước lượng khoảng với độ tin cậy 0,90.

d) Ước lượng tỷ lệ gia đình có thu nhập $10 < X \leq 20$ và $X > 20$.

Các tỷ lệ trên thấp nhất là bao nhiêu? (cho độ tin cậy là 0,95).

3. Trong lần điều tra dân số năm 1987, dân số của tỉnh Z là: 515480 trẻ em dưới 15 tuổi, 200685 thanh thiếu niên tuổi từ 15 đến dưới 18 tuổi; 1046830 người tuổi từ 18 đến 60 và 135270 người trên 60 tuổi. Trong lần điều tra dân số năm 1999 dân số của tỉnh Z là: 373179 trẻ em dưới 15 tuổi, 342949 thanh thiếu niên tuổi từ 15 đến dưới 18; 1532092 người từ 18 đến 60 tuổi và 266738 người trên 60 tuổi. Hãy biểu diễn phân bố dân số tỉnh Z theo lứa tuổi trong hai lần điều tra dân số bằng biểu đồ tần suất, biểu đồ chữ nhật. Nhận xét gì về sự phát triển dân số của tỉnh Z sau 12 năm?

4. Gọi X là điểm thi môn Thống kê xã hội của sinh viên nhóm ngành VII. Giả sử $X \simeq N(\mu, \sigma^2)$. Tổng kết điểm thi của 1 lớp gồm 150 sinh viên, ta tính được $\bar{X} = 7,5$; $s = 1,5$. Giả sử kết quả thi của lớp này phản ánh được kết quả thi của cả khóa (gồm 2500 sinh viên). Từ kết quả nhận được có thể kết luận điểm thi trung bình môn Thống kê của cả khóa thuộc khoảng nào với độ tin cậy 95% ? Có thể nói điểm thi trung

bình của cả khóa cao hơn 7 được không? (Cho mức ý nghĩa $\alpha = 0,05$).

5. Gọi X là chỉ số thông minh (IQ) của học sinh lứa tuổi 12 – 15.

Giả sử X có phân phối chuẩn N (μ; 9). Đo IQ ở 50 em ta có các số liệu sau:

Khoảng giá trị X	< 75	[75;78)	[78;81)	[81;84)	[84;87)	[87;90)	[90;93)	≥ 93
Số học sinh	1	2	8	9	12	10	7	1

a) Hãy chỉ ra ước lượng cho chỉ số IQ trung bình.

b) Từ kết quả trên có thể nói chỉ số IQ trung bình đang xét là trên 84 không (với mức ý nghĩa $\alpha = 0,05$)?

c) Với xác suất 90% có thể nói chỉ số IQ trung bình thấp nhất là bao nhiêu?

d) Với độ tin cậy 98% có thể nói tỷ lệ học sinh có IQ trên 84 nằm trong khoảng nào?

6. Điều tra ngẫu nhiên 100 sinh viên được biết có 30 trường hợp chọn ngành nghề học là theo sự gợi ý của bố mẹ.

a) Hãy cho ước lượng về tỷ lệ (p) sinh viên chọn ngành học theo sự gợi ý của bố mẹ.

b) Với xác suất 90% có thể nói tỷ lệ p cao nhất là bao nhiêu?

7. Gọi p là tỷ lệ các vụ tai nạn giao thông đường bộ do thanh thiếu niên gây ra ở thành phố Z. Theo tổng kết của năm trước thì tỷ lệ p là 62%. Sau một năm thực hiện Nghị định 36/CP, các trường học đưa giáo dục về luật giao thông

vào dạy cho học sinh, trong số 1500 vụ tai nạn có 720 vụ là do thanh thiếu niên gây ra. Hãy trả lời các câu hỏi sau trong hai tình huống: số liệu trên là mẫu đại diện và số liệu trên là số liệu tổng kết đầy đủ.

a) Cho ước lượng tỷ lệ p sau một năm triển khai Nghị định 36/CP.

b) Từ kết quả nhận được có thể nói tỷ lệ p đã giảm so với năm trước hay không? Thậm chí có kết luận p nhỏ hơn 50% hay không (xét với mức ý nghĩa $\alpha = 0,05$).

c) Với xác suất 90% có thể nói tỷ lệ p cao nhất là bao nhiêu?

8. Trong 6 tháng đầu năm tỷ lệ tội phạm ở thanh thiếu niên là 40%. Tổng kết 6 tháng cuối năm (sau khi có Nghị định 87/CP) vẫn ở các địa phương trên cho thấy trong số 644 vụ phạm tội có 161 vụ do thanh thiếu niên gây ra.

a) Hãy chỉ ra ước lượng điểm cho tỷ lệ tội phạm ở thanh thiếu niên sau khi có Nghị định 87/CP (ký hiệu tỷ lệ đang xét là p).

b) Với độ tin cậy 90% có thể nói tỷ lệ p nằm trong khoảng nào.

c) Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận tỷ lệ tội phạm ở thanh thiếu niên trong 6 tháng cuối năm đã giảm so với 6 tháng đầu năm hay không?

d) Nếu số liệu tổng kết trên không phải là mẫu đại diện mà là số liệu tổng kết đầy đủ và chính xác thì ba câu hỏi trên được thay bởi câu hỏi nào và kết luận ra sao?

9. Thống kê thời gian tự học ở nhà trong một tuần lễ của 70 sinh viên năm thứ I khoa A, ta có số liệu mẫu sau:

Thời gian T (giờ)	< 5	[5;10)	[10;15)	[15;20)	[20;25)	[25;30)	[30;35)	[35;40)	≥ 40
Số sinh viên	3	7	10	18	12	8	5	4	3

Coi như thời gian T tuân theo luật phân phối chuẩn.

a) Hãy ước lượng thời gian tự học ở nhà trung bình của một sinh viên khoa A.

b) Từ ước lượng nhận được có thể kết luận thời gian tự học ở nhà trung bình của một sinh viên khoa A là trên 20 giờ/1 tuần hay không? (Xét với mức ý nghĩa $\alpha = 10\%$).

c) Với xác suất 95% có thể nói thời gian tự học ở nhà trung bình của một sinh viên khoa A cao nhất là bao nhiêu? Khả năng đúng, sai của kết luận trên là bao nhiêu?

d) Hãy ước lượng tỷ lệ sinh viên có trách nhiệm với việc học tập của mình ($T \geq 30$). Với độ tin cậy 90% tỷ lệ này nằm trong khoảng nào? Thấp nhất là bao nhiêu?

10. Một loại thuốc chữa bệnh được nhà sản xuất khẳng định xác suất khỏi bệnh khi dùng thuốc này là 80%. Đưa thuốc ra sử dụng ở một bệnh viện. Theo dõi trong 150 người dùng, thấy có 110 người khỏi. Với mức ý nghĩa $\alpha = 0{,}05$ có thể chấp nhận ý kiến của nhà sản xuất hay không? Nếu không thì xác suất trên cao hay thấp hơn 80%?

11. Một công ty dự định mở một siêu thị tại khu dân cư M. Để đánh giá khả năng mua hàng của khách hàng khu vực này, người ta đã điều tra ngẫu nhiên thu nhập trong một tháng của 100 hộ và thu được các số liệu sau:

Thu nhập X (triệu đồng)	4	4,5	5	5,5	6	6,5	7	7,5
Số hộ	5	10	15	20	29	10	6	5

a) Theo bộ phận tiếp thị thì chỉ nên mở siêu thị tại M nếu thu nhập trung bình của các hộ tối thiểu là 5,5 triệu đồng. Với mức ý nghĩa 5% hãy cho biết có nên mở siêu thị tại khu dân cư M hay không?

b) Với độ tin cậy 95%, hãy tìm khoảng tin cậy cho tỷ lệ các hộ có thu nhập ≥ 5,5 triệu đồng/tháng.

c) Điều tra ngẫu nhiên 100 hộ ở khu dân cư N, ta tính được thu nhập bình quân tháng là 5,8 triệu đồng và s = 1,1 triệu đồng. Có thể kết luận thu nhập bình quân ở khu dân cư N cao hơn ở khu M hay không? (Cho mức ý nghĩa $\alpha = 0,05$).

12. Tham gia dự án trồng 5 triệu ha rừng, nhân dân xã Z đã được giao trồng 100 ha rừng bằng hai loại cây: bồ đề và keo lai. Sau hai tuần lễ người ta lấy mẫu ngẫu nhiên và kết quả nhận được như sau: trong số 200 cây bồ đề thì có 175 cây sống, trong số 150 cây keo lai thì có 110 cây sống. Với mức ý nghĩa $\alpha = 0,10$ có thể kết luận tỷ lệ cây bồ đề sống cao hơn cây keo lai hay không?

13. Để thăm dò ý kiến của người tiêu dùng, bộ phận tiếp thị của hãng đã điều tra ngẫu nhiên và được số liệu mẫu sau:

Hỏi 700 nam giới thì có 280 người thích dùng dầu gội đầu Dr.Men và 265 người thích dùng dầu gội Clear.

Với mức ý nghĩa $\alpha = 0,02$ có thể kết luận người tiêu dùng không thích dùng Clear bằng Dr.Men hay không?

14. Theo các nhà nhân trắc học, hiện nay chiều cao của các em lứa tuổi 14 – 15 ở thành thị cao hơn ở nông thôn từ 2cm đến 3cm.

Ở nông thôn, lấy ra một mẫu với $n_1 = 100$, tính được $\bar{X} = 140,5$cm

Ở thành thị ta lấy ra một mẫu với $n_2 = 80$, tính được $\bar{Y} = 142,9cm$

Giả sử DX = DY = 9.

Với mức ý nghĩa $\alpha = 0,10$ có thể khẳng định chiều cao trung bình của các em lứa tuổi 14 – 15 ở thành thị cao hơn ở nông thôn hay không?

15. Để đánh giá chất lượng dạy và học môn tiếng Anh ở hai trường THCS A và B, người ta lấy ra hai mẫu đại diện nhân đợt kiểm tra chất lượng chung của toàn Quận. Kết quả như sau:

Điểm thi	2	3	4	5	6	7	8	9	10
Số học sinh trường A	1	5	10	20	25	18	12	6	3
Số học sinh trường B	0	3	7	12	20	28	15	8	7

Giả sử điểm thi tuân theo luật phân phối chuẩn, DX = DY.

Với mức ý nghĩa $\alpha = 0,05$.

a) Hãy so sánh kết quả thi trung bình của hai trường.

b) Tỷ lệ điểm khá giỏi của hai trường có như nhau không?

c) Tỷ lệ điểm trung bình của hai trường có hơn, kém nhau hay không?

d) Hãy ước lượng tỷ lệ điểm giỏi của hai trường. Sự khác nhau giữa hai ước lượng này là ngẫu nhiên hay bản chất?

e) Qua các kết luận nhận được ở trên, có thể nói gì về chất lượng dạy và học môn tiếng Anh của 2 trường THCS A và B ?

16. Để xem doanh số bán hàng trong một năm của một công ty bảo hiểm giữa những người đã tốt nghiệp trường

thương nghiệp và những người không qua trường thương nghiệp có khác nhau không, người ta lấy ra hai mẫu ngẫu nhiên với kết quả như sau (đơn vị: triệu đồng)

Nhóm I (x_i):	300	310	290	260	280	320	310	280	250	300
	350	320	320	290	270	280	310	290	300	290
Nhóm II (y_j):	250	220	280	200	190	220	270	250	260	
	180	200	220	240	220	260				

Với mức ý nghĩa $\alpha = 0,10$ thử xem doanh số bán hàng của hai nhóm trên có như nhau không hay nhóm nào cao hơn?

17. Có ý kiến cho rằng tăng giá điện sẽ làm cho người dùng tiết kiệm điện hơn. Thống kê mức tiêu thụ điện ở 18 hộ gia đình (thuộc loại có mức sống ổn định). Gọi X và Y là số kW/giờ tiêu thụ trong một tháng trước và sau khi tăng giá điện của mỗi gia đình (ảnh hưởng thời tiết của 2 tháng này là không đáng kể). Ta có các số liệu sau:

x_i:	300	310	300	290	305	315	312	306	295	288
y_i:	305	305	290	295	315	311	309	308	299	285
x_i:	330	340	320	315	303	320	287	270	295	338
y_i:	340	328	316	307	295	320	290	265	295	332

Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận được gì về ý kiến nêu trên? Rút ra ý nghĩa thực tiễn gì từ kết luận nhận được?

18. Kết quả thi học kỳ, gồm 5 môn tự nhiên và 5 môn xã hội, của 18 học sinh lớp chuyên được chọn một cách ngẫu nhiên, được cho tương ứng như sau:

Điểm 5 môn tự nhiên (x_i)	45	48	42	38	39	43	46	40	47
Điểm 5 môn xã hội (y_i)	43	45	38	35	33	36	41	38	39
Điểm 5 môn tự nhiên (x_i)	46	44	43	35	34	32	33	35	39
Điểm 5 môn xã hội (y_i)	36	35	38	30	36	34	33	36	34

Với mức ý nghĩa $\alpha = 0,02$ liệu có thể kết luận học sinh lớp chuyên này học lệch về các môn tự nhiên hơn hay không?

19. Có ý kiến cho rằng trong các cuộc thi "Đi tìm triệu phú" trên VTV3, những người tham gia chơi là nữ thường đạt kết quả tốt hơn nam. Điều tra lại kết quả nhận được của 18 nữ và 15 nam ta có các số liệu sau (đơn vị: triệu đồng)

Kết quả của nữ:	5	1	10	7	7	3,2	5	1	7	10
	15	7	15	10	7	10	1	5.		
Kết quả của nam:	7	1	1	5	10	1	5	15	1	1
	5	7	1	10	3,2.					

Với mức ý nghĩa $\alpha = 0,10$ hỏi ý kiến trên có đúng hay không?

20. Trong báo cáo đánh giá chất lượng học môn Lịch sử lớp 10 ở một trường nọ, có 10% giỏi, 25% khá, 50% trung bình, 15% kém. Trong lần thi cuối học kỳ, người ta lấy ngẫu nhiên 100 bài và tổ chức chấm lại thấy có 2 bài đạt điểm giỏi, 10 bài đạt điểm khá, 48 bài điểm trung bình, 40 bài điểm kém.

a) Với số liệu có được, có thể chấp nhận báo cáo của nhà trường hay không (xét với $\alpha = 0,05$)?

b) Có thể chấp nhận được tỷ lệ điểm nào trong các tỷ lệ mà trường báo cáo hay không (với mức ý nghĩa $\alpha = 0,05$)?

c) Với xác suất 0,95 tỷ lệ điểm kém thấp nhất là bao nhiêu %?

21. Điều tra số vụ tai nạn giao thông do môtô (xe máy) gây ra ở ba địa phương ta có các số liệu sau:

Ở địa phương A: trong số 120 vụ tai nạn thì có 70 vụ do môtô (xe máy).

Ở địa phương B: trong số 90 vụ tai nạn thì có 54 vụ do môtô (xe máy).

Ở địa phương C: trong số 140 vụ tai nạn thì có 100 vụ do môtô (xe máy).

Với mức ý nghĩa $\alpha = 0,05$ hỏi tỷ lệ tai nạn do môtô (xe máy) gây ra ở ba địa phương trên có như nhau không?

22. Để nghiên cứu xem ngoài năng khiếu thì kết quả của việc học ngoại ngữ và các môn tự nhiên có ảnh hưởng lẫn nhau hay không, người ta điều tra ngẫu nhiên một số trường hợp. Kết quả ta được các số liệu sau:

Ngoại ngữ / Tự nhiên	Khá + Giỏi	Trung bình	Kém
Khá + Giỏi	45	8	0
Trung bình	20	50	2
Kém	4	10	3

a) Với mức ý nghĩa $\alpha = 0,10$ hãy kiểm tra xem kết quả học môn ngoại ngữ và các môn tự nhiên có mối liên hệ gì với nhau hay không hay hoàn toàn độc lập?

b) Từ kết luận nhận được, anh (chị) thấy có phù hợp với thực tiễn không? Giải thích.

23. Kết quả thi môn học chung của sinh viên ba khoa được thể hiện qua mẫu dưới đây:

Kết quả / Khoa	Giỏi	Khá	Trung bình	Kém
A	15	18	20	3
B	6	12	25	7
C	8	15	22	5

Với mức ý nghĩa $\alpha = 0,05$ hãy kiểm tra xem kết quả thi môn học chung đó có phụ thuộc vào khoa hay không?

24. Nghiên cứu tình trạng hôn nhân (TTHN) trước ngày cưới của 542 cặp vợ chồng, ta có bảng số liệu sau:

TTHN của vợ / TTHN của chồng	Chưa kết hôn lần nào	Ly hôn	Góa
Chưa kết hôn lần nào	180	34	36
Ly hôn	58	76	54
Góa	43	34	27

Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận tình trạng hôn nhân của vợ và của chồng trước ngày cưới là độc lập với nhau hay không?

25. Một công ty muốn xác định xem sự nghỉ việc có liên quan đến tuổi tác hay không, họ đã tiến hành điều tra mẫu và cho kết quả như sau (xem [2]):

Tuổi / Lý do	Dưới 30	Từ 30 đến 50	Trên 50
Ốm đau	40	28	52
Lý do khác	20	36	24

Vậy có sự phụ thuộc giữa việc nghỉ và tuổi tác hay không? Hoặc tỷ lệ nghỉ việc do ốm đau ở ba lớp tuổi có như nhau không? (với mức ý nghĩa $\alpha = 0,10$).

26. Điều tra sức khoẻ trẻ em ở tỉnh A người ta đưa ra kết quả sau: 11% trẻ bị suy dinh dưỡng nặng; 33% trẻ bị suy dinh dưỡng. Ở tỉnh B, khi điều tra 3000 trẻ, thấy có 270 trẻ suy dinh dưỡng nặng; 830 trẻ bị suy dinh dưỡng. Với mức ý nghĩa $\alpha = 0,10$:

a) Có thể nói tình trạng sức khoẻ của trẻ ở tỉnh B cũng như ở tỉnh A hay không?

b) Tỷ lệ trẻ bị suy dinh dưỡng nặng ở tỉnh B có khác tỉnh A hay không? Nếu khác thì ở tỉnh nào cao hơn?

27. Tai nạn giao thông đang là vấn đề bức xúc của toàn xã hội. Trong 9 tháng đầu năm 2002 cả nước đã xảy ra 21312 vụ, trong đó có 9584 người bị chết, 23981 người bị thương. Nhìn chung, bình quân mỗi ngày cả nước có 30 người bị chết, mỗi tháng có khoảng 1000 người bị chết vì tai nạn giao thông. Thống kê trong vài năm về tỷ lệ tăng xe máy (X%) và tỷ lệ gia tăng số người chết vì tai nạn giao thông (Y%), ta có các số liệu sau:

Năm	1993	1995	1997	2000	2001
Tỷ lệ tăng xe máy (X%)	42,4	13,0	6	15,7	29,6
Tỷ lệ tăng số người chết (Y%)	43,0	7,5	4,8	12,4	39,7

a) Có nhận xét gì về sự gia tăng tỷ lệ xe máy và tỷ lệ người chết do tai nạn giao thông qua số liệu đã cho?

b) Hãy tính hệ số tương quan giữa X và Y. Kết quả nhận được có giúp anh (chị) chứng minh nhận xét trên hay không? Giải thích cụ thể.

28. Để nghiên cứu mối quan hệ giữa số người tiêm chích ma túy (X) với số người nhiễm HIV (Y), giả sử người ta thống kê ở 10 địa phương (tỉnh, thành phố) trọng điểm trong cả nước, có các số liệu sau (không phải là số liệu điều tra thực):

x_i	100	150	200	250	300	300	350
y_i	60	90	120	150	150	180	180
Số địa phương (m_i)	1	1	1	2	1	3	1

a) Tính hệ số tương quan mẫu r.

b) Có nhận xét gì về sự phụ thuộc giữa số người nhiễm HIV và số người tiêm chích ma túy?

c) Xây dựng đường hồi quy tuyến tính thực nghiệm của y theo x?

d) Đường hồi quy nhận được có thể dùng để làm gì? Vì sao? Cho ví dụ.

e) Ước lượng sai số bình phương trung bình mắc phải khi dùng đường hồi quy trên.

29. Thống kê số người tham gia đi du lịch, nghỉ mát của cơ quan nọ ta có các số liệu sau (X là mức thu nhập thêm / tháng (triệu đồng), Y là tỷ lệ người tham gia (%)).

x_i	< 0,5	[0,5–0,8)	[0,8–1,1)	[1,1–1,4)	[1,4–1,7)	[1,7–2,0)
y_i	5	15	25	45	55	65

a) Tính hệ số tương quan mẫu r.

b) Nhận xét gì về sự phụ thuộc giữa tỷ lệ người tham gia và mức thu nhập thêm?

c) Xây dựng đường hồi quy tuyến tính thực nghiệm của y theo x.

30. Trung tâm Thông tin thư viện – Đại học Quốc gia Hà Nội đã phát hành "Bản tin điện tử" hàng tháng. Ban biên tập đã thăm dò ý kiến của một số cán bộ ở 3 khoa về chất lượng của bản tin. Kết quả được cho như sau (tình huống và số liệu không phải là thực, chỉ có tính chất ví dụ):

Khoa / Đánh giá	A	B	C
Chất lượng tốt	30	20	25
Chất lượng trung bình	15	15	12
Chất lượng yếu hoặc không quan tâm	5	3	8

a) Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận kết quả đánh giá chất lượng bản tin phụ thuộc vào khoa được hỏi hay không?

b) Tỷ lệ chung cán bộ đánh giá bản tin có chất lượng tốt là 55%. Với mức ý nghĩa $\alpha = 0,05$ có thể nói cán bộ khoa A có tỷ lệ đánh giá chất lượng tốt cao hơn mức đánh giá chung hay không?

c) Trong số 133 cán bộ được hỏi có 16 người đánh giá chất lượng yếu hoặc họ không quan tâm. Với độ tin cậy 90% có thể nói tỷ lệ trên cao nhất là bao nhiêu % ?

d) Với mức ý nghĩa $\alpha = 0,05$ tỷ lệ đánh giá chất lượng tốt ở 3 khoa có như nhau không?

31. Quan sát 10 gia đình ở một vùng có nghề phụ thủ công theo hai tiêu thức: Số trẻ em dưới 16 tuổi, nhưng lớn hơn 7 tuổi (X) và thu nhập thêm bằng nghề phụ (Y: nghìn đồng) người ta thu được các số liệu sau:

x	3	5	2	4	4	4	6	1	3	3
y	58	89	72	71	68	64	98	49	59	62

a) Hãy khảo sát mối tương quan giữa hai tiêu thức trên.

b) Xây dựng đường hồi quy tuyến tính thực nghiệm của thu nhập theo số trẻ.

c) Đường hồi quy nhận được có thể dùng làm gì? Vì sao?

d) Ước lượng sai số bình phương trung bình.

HƯỚNG DẪN GIẢI VÀ ĐÁP SỐ

Chương I
MỘT SỐ KHÁI NIỆM VÀ KẾT QUẢ CƠ BẢN CỦA XÁC SUẤT

1. Phép thử ở đây là chọn 3 trong 10. Chọn 1 lần. Chọn theo nghĩa của tổ hợp. Số biến cố sơ cấp là $C_{10}^3 = 120$.

a) $\dfrac{C_3^3 C_7^0}{C_{10}^3} = \dfrac{1}{120} = 0,0083$

b) $\dfrac{C_3^2 C_7^1}{C_{10}^3} = \dfrac{21}{120} = 0,175$

c) $\dfrac{3.21}{120} = 0,525$

d) $\dfrac{35}{120} = 0,275$

e) Ta dự đoán trong 3 người chọn ra sẽ có 1 đại biểu dân tộc ít người. Đoán như vậy khả năng đúng (0.525) cao hơn 3 tình huống còn lại.

2. Phép thử là lấy 1 lần.

Số biến cố sơ cấp là $C_{52}^6 = 20.358.520$

a) $\dfrac{C_{26}^4 . C_{26}^2}{C_{52}^6} = \dfrac{14950}{20.358.520} = 0,00073$

b) $\dfrac{C_3^1 C_{13}^2 C_{13}^1 C_{13}^2}{C_{52}^6} = \dfrac{1028196}{20.358.520} = 0,0505$

c) $\dfrac{C_4^1 C_4^3 C_4^2}{C_{52}^6} = \dfrac{96}{20.358.520} = 0,0000047$

3. Phép thử là lấy 1 lần.

Số biến cố sơ cấp là $C_{15}^3 = 455$

a) $\frac{C_5^1 C_6^1 C_4^1}{C_{15}^3} = \frac{120}{455} = 0,2637$ b) $\frac{C_5^2 C_6^1}{C_{15}^3} = \frac{60}{455} = 0,1319$

c) $\frac{C_6^3}{C_{15}^3} = \frac{20}{455} = 0,0440$ d) $\frac{C_4^2 C_5^1}{C_{15}^3} = \frac{30}{455} = 0,0659$

4.

a) Số biến cố sơ cấp = 6.6 = 36

b) Đếm ra ta thấy có 1 khả năng X = 2 ; 2 khả năng X = 3; 3 khả năng X = 4. Do đó: $P(1 \le X < 5) = \frac{1+2+3}{36} = \frac{1}{6}$

Tương tự: $P(10 \le X < 15) = \frac{3+2+1}{36} = \frac{1}{6}$;

$P(12 < X < 15) = \frac{0}{36} = 0$

5. Phép thử là: 8 lần ghi tên (vì có 8 người). Mỗi lần ghi có 3 cách chọn. Số biến cố sơ cấp là $3^8 = 6561$.

a) $\frac{1}{3^8} = 0,00015$ b) $\frac{3.1.1...1}{3^8} = 0,00046$

c) $\frac{C_8^5 C_3^2 C_1^1}{3^8} = \frac{168}{6561} = 0,0256$ d) $\frac{3.1.3^6}{3^8} = \frac{1}{3} = 0,3333$

(Số thuận lợi cho c): Chọn 5 trong 8 người đưa vào cắm hoa (có C_8^5 cách chọn). Sau đó chọn 2 trong 3 người còn lại đưa vào thi nấu ăn (có C_3^2 cách chọn). Cuối cùng chỉ còn C_1^1 cách chọn người vào dancing. Số thuận lợi cho d): A có 3 cách chọn, B chỉ có 1 cách; 6 người còn lại, không có ràng buộc gì, mỗi người vẫn có 3 cách chọn).

6. Z = {Sinh viên Z thi đạt}, T = {Sinh viên T thi đạt}

Z, T là 2 biến cố độc lập, không xung khắc.

a) P(Z T) = 0,8075

b) P(Z $\cup$ T) = 0,9925

c) P($\overline{Z}$ $\overline{T}$) = 0,0075

7. Dùng mô hình tính xác suất bằng công thức xác suất của tổng và của tích.

a) 0,760 b) 0,990 c) 0,190 d) 0,230

8. Xét 100 vụ ly hôn ta có 100 phép thử Bernoulli với biến cố A = {Ly hôn do ngoại tình} ; p = 0,30.

a) $P_{100}(40 ; 0,30) = C_{100}^{40}.0,3^{40}.0,7^{60}$

b) $P_{100}(50 ; 0,70) = C_{100}^{50}.0,7^{50}.0,3^{50}$

c) $P_{100}(m \geq 1; 0,30) = 1 - P_{100}(0; 0,30)$

$= 1 - C_{100}^{0}.0,3^{0}.0,7^{100} = 1 - 0,7^{100}$

d) np + p – 1 = 100.0,3 + 0,3 – 1 = 29,3 $\rightarrow$ m_0 = 30.

Trong 100 vụ ly hôn thì 30 vụ ly hôn do ngoại tình là có khả năng xảy ra cao nhất.

9. 1240 cặp vợ chồng đăng ký ở phường trong 10 năm qua, như bài 8, có thể xem như ta đã lấy ra 1240 trường hợp từ tập các cặp vợ chồng đã đăng ký trong các năm qua ở cả thành phố, cũng như trong cả nước (vì các tỷ lệ đã cho trong bài là tỷ lệ chung của cả nước, chứ không phải chỉ của phường này). Do đó, ta có 1240 phép thử Bernoulli với $p = \frac{1}{8}$.

a) $np + p - 1 = 1240.\frac{1}{8} + \frac{1}{8} - 1 = 154,125 \rightarrow m_0 = 155.$

Số cặp vô sinh có khả năng xảy ra cao nhất là 155 cặp.

Xác suất tương ứng là $C_{1240}^{155}\left(\frac{1}{8}\right)^{155}\left(\frac{7}{8}\right)^{1085}$.

b) Xét 155 cặp vô sinh, ta lại có 155 phép thử Bernoulli với p = 0,40.

Cho nên số cặp vô sinh do chồng có khả năng nhất là 62 cặp (vì $np + p - 1 = 155.0,4 + 0,4 - 1 = 61,4 \rightarrow m_0 = 62$).

c) $P_{155}(m \leq 2 ; 0,6) = P_{155}(0 ; 0,6) + P_{155}(1 ; 0,6) + P_{155}(2 ; 0,6).$

$= C_{155}^{0}.(0,4)^{155} + C_{155}^{1}.0,6.0,4^{154} + C_{155}^{2}.0,6^2.0,4^{153}$

$= 0,4^{153}(0,4^2 + 155.0,6.0,4 + 11935.0,6^2) = 4333,96.(0,4)^{153}$

(Vì $m \leq 2$, tức là m = 0 hoặc m = 1 hoặc m = 2. Ba biến cố này xung khắc (do nếu m lấy giá trị 1 thì m không lấy giá trị 0 hoặc 2 nữa,...)).

10.

a) Ta có:

X	1	2	3	4	5
P(X = m)	0,7	0,3.0,7	$0,3^2.0,7$	$0,3^3.0,7$	$0,3^4$
	0,7	0,21	0,063	0,0189	0,0081

(X = 2) = (Viên I trượt) (Viên II trúng)

(X = 5) = (4 viên đầu đều trượt) (Viên thứ V hoặc trúng hoặc trượt)

Hoặc:

P(X = 5) = 1 – P(X = 1) – P(X = 2) – P(X = 3) – P(X = 4).

b)

$$F(x) = \begin{cases} 0 & \text{nếu } x \le 1 \\ 0,7 & \text{nếu } 1 < x \le 2 \\ 0,91 & \text{nếu } 2 < x \le 3 \\ 0,973 & \text{nếu } 3 < x \le 4 \\ 0,9919 & \text{nếu } 4 < x \le 5 \\ 1,0 & \text{nếu } 5 < x \end{cases}$$

c) $EX = 1,4251$

$EX^2 = 2,6119$

$DX = 0,5810$

$ModX = 1$

$P(1 \le X \le 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0,973.$

$P(2 < X \le 5,4) = P(X = 3) + P(X = 4) + P(X = 5) = 0,090.$

11.

a) Ta có:

$Z = 3X$	-6	-3	0	3
$P(Z = z_i)$	0,1	0,3	0,4	0,2

$Z = X^2$	4	1	0
$P(Z = z_i)$	0,1	0,5	0,4

$Z = X + Y$	-3	-2	-1	0	1	2
$P(Z = z_i)$	0,03	0,09	0,19	0,27	0,28	0,14

b) $E(X + Y) = 0,10$

$D(X - Y) = DX + DY = 0,81 + 0,84 = 1,65.$

(Hoặc $D(X - Y) = DX + DY = D(X + Y) = E(X + Y)^2 - 0,1^2$).

$Mod(X + Y) = 1$.

$P(-2 < X + Y \leq 2,1) = 0,88$.

12.

a) Khảo sát ngẫu nhiên 50 trẻ ở vùng này chính là 50 phép thử Bernoulli với $p = 0,358$; nên $X = B(50 ; 0,358)$.

b) Về trung bình sẽ có 17,9 trẻ suy dinh dưỡng

c) $np + p - 1 = 50.0,358 + 0,358 - 1 = 17,258 \rightarrow m_0 = 18$.

Có 18 trẻ suy dinh dưỡng là khả năng xảy ra cao nhất.

d) Có thể xảy ra tình huống cả 50 trẻ đều bị suy dinh dưỡng, nhưng khả năng xảy ra là khá nhỏ. Còn 18 trẻ suy dinh dưỡng sẽ có khả năng xảy ra cao nhất.

e) $P_{50} (m \geq 1 ; 0,642) = 1 - P_{50} (0 ; 0,642) = 1 - 0,358^{50}$.

13.

a) 120 cháu của nhà trẻ mẫu giáo có thể xem như được chọn từ tập các cháu dưới 5 tuổi, mà tập này có tỷ lệ thừa cân là 1,7%. Vì vậy, ta có 120 phép thử Bernoulli với $p = 0,017$.

Số trẻ thừa cân có phân phối $B(120 ; 0,017)$.

b) $np + p - 1 = 1,057 \rightarrow m_0 = 2$

Có 2 cháu thừa cân là tình huống xảy ra cao nhất.

c) Số tìm được không phải là số trung bình, vì số trung bình là 2,04

d) Mặc dù số trung bình là 2,04; số có khả năng nhất là 2 nhưng tình huống có trên 10 cháu bị thừa cân vẫn có khả năng xảy ra, chỉ có điều là xác suất xảy ra sẽ nhỏ.

14.

X	0	1	2	3	4
P(X = m)	$\frac{C_4^0 C_8^5}{C_{12}^5}$	$\frac{C_4^1 C_8^4}{C_{12}^5}$	$\frac{C_4^2 C_8^3}{C_{12}^5}$	$\frac{C_4^3 C_8^2}{C_{12}^5}$	$\frac{C_4^4 C_8^1}{C_{12}^5}$
	$\frac{56}{792}$	$\frac{280}{792}$	$\frac{336}{792}$	$\frac{112}{792}$	$\frac{8}{792}$
	0,071	0,353	0,424	0,141	0,111

Các câu còn lại khá quen thuộc. Bạn đọc tự giải.

15.

a) Giả thiết Y = N(140 ; 9) cho biết rằng chiều cao trung bình của các em lứa tuổi 13 là 140cm. Chiều cao của các em sẽ dao động quanh 140 với độ lệch tiêu chuẩn là 3cm, tức là 68,26% các em sẽ có chiều cao thuộc khoảng (137cm ; 143cm); 95,45% các em sẽ có chiều cao thuộc khoảng (134cm; 146cm) và 99,73% các em sẽ có chiều cao thuộc khoảng (131cm; 149cm) (theo quy tắc kσ).

b) Tỷ lệ các em có chiều cao thuộc khoảng (135 ; 150) chính là: P(135 < Y < 150), nghĩa là ta chọn ngẫu nhiên một em 13 tuổi, xác suất chọn được em có chiều cao thuộc khoảng (135 ; 150) chính là P(135 < Y < 150). Theo công thức (I.7) mục I.9.3 ta có:

$$P(135 < Y < 150) = \Phi\left(\frac{10}{3}\right) - \Phi\left(-\frac{5}{3}\right) = 0{,}9522.$$

(X liên tục nên khoảng trên kín hay hở ở 2 đầu mút, xác suất đều như nhau)

c) 50 em trong lớp lại được coi như 50 phép thử Bernoulli. Với A = (135 < Y < 150) và xác suất p = 0,9522 ta có:

i) $Z \simeq B(50 ; 0{,}9522)$

$\leftrightarrow P(Z = m) = C_{50}^{m}.0{,}9522.0{,}0478^{50-m}$

$m = \overline{0 ; 50}$

ii) Về trung bình, trong 50 em sẽ có 47,61 em có chiều cao từ 135cm đến 150cm.

iii) Ta dự đoán trong 50 em sẽ có 48 em có chiều cao từ 135cm đến 150cm. Vì xác suất xảy ra trường hợp này là cao nhất. Dự đoán như vậy sẽ có khả năng đúng lớn nhất.

Chương II

THỐNG KÊ XÃ HỘI HỌC

1. b) Mẫu thu gọn nhận được:

x_i:	12	16	25	26	29	31	32	33	34	36	37	38	39
m_i:	1	1	1	1	1	1	1	2	1	2	2	3	2
x_i:	40	41	43	44	45	46	47	48	49	50	51	53	54
m_i:	2	4	2	5	7	1	3	2	2	2	4	2	1

x_i:	55	56	57	58	59	61	62	63	64	65	66	68	69
m_i:	5	2	3	2	2	3	1	1	2	1	3	2	1

x_i:	71	72	73	76	77	77	78	80	82	83	92
m_i:	2	2	2	1	2	2	2	1	2	2	2

Điểm $x_i = 45$ có tần số cao nhất.

d) Từ số liệu a) hoặc b) có: MedX = 51

e) Từ số liệu a) hoặc b) có:

$$Q_1 = \frac{1}{2}(41 + 43) = 42; \qquad Q_3 = \frac{1}{2}(63 + 64) = 63,5.$$

f) $\bar{X} = 52,67$; $s = 15,71$; $\hat{s} = 15,79$.

g) Đây là ước lượng khoảng EX. Dùng khoảng c) ứng với trường hợp phương sai chưa biết, không có giả thiết X chuẩn, nhưng n lớn: u(0,025) = 1,96.

Với độ tin cậy 95%: EX ∈ (49,575 ; 55,765)

h) $p^* = \frac{6}{100} = 0,06$

Tỷ lệ sinh viên đạt điểm giỏi là 6%.

Với độ tin cậy 0,90, tỷ lệ sinh viên đạt điểm giỏi thuộc khoảng (2,1% ; 9,9%).

2. a) $\text{Med}X = \frac{\frac{215}{2} - 102}{19}(16 - 15) + 15 = 15,2895$

$Q_1 = \frac{\frac{215}{4} - 49}{15}(13 - 12) + 12 = 12,3167$

$Q_3 = \frac{3.\frac{215}{4} - 155}{14}(19 - 18) + 18 = 18,4464$

b) Khoảng cuối cùng > 24 ta thay bởi điểm 24,5 (như ví dụ II.6 đã xử lý).

$\bar{X} = 15,4163$; $s = 4,2768$; $\hat{s} = 4,2867$; $\hat{s}^2 = 18,3761$

c) Mức thu nhập trung bình hằng năm của các gia đình là 15.416.300 đồng

Với độ tin cậy 90%, μ ∈ (14.945.600 ; 15.887.000)

(Tra bảng ta lấy $t_{214}(0,05) \approx 1,61$)

d) Gọi $p_1 = P(10 < X \leq 20)$ và $p_2 = P(X > 20)$.

Ta có: $p_1^* = \frac{158}{215} = 0,735$; $p_2^* = \frac{34}{215} = 0,158$

$1 - \alpha = 0,95 \rightarrow \alpha = 0,05 \rightarrow u(0,025) = 1,96$

Với xác suất 95% các tỷ lệ trên thấp nhất là:

$p^*_{1\min} = 0,676 = 67,6\%.$

$p^*_{2\min} = 0,109 = 10,9\%$

3. Số liệu đã cho là số liệu đầy đủ, chính xác, không phải là số liệu mẫu đại diện.

Tỷ lệ phần trăm số dân ở các loại tuổi tương ứng là:

Năm 1987	27,15%	10,57%	55,15%	7,13%
Năm 1999	14,69%	13,50%	60,31%	10,50%

Biểu diễn trên biểu đồ tần suất và biểu đồ chữ nhật như ví dụ II.2 đã làm.

Nhìn vào biểu đồ ta thấy tỷ lệ trẻ em dưới 15 tuổi giảm xuống gần nửa, chứng tỏ tỷ lệ sinh đẻ đã giảm xuống. Tỷ lệ người cao tuổi tăng lên, chứng tỏ số người quá tuổi lao động, số người hưởng hưu trí tăng lên, phần nào cũng chứng tỏ tuổi thọ cũng được tăng lên. Tỷ lệ người ở lứa tuổi từ 15 đến 60 tăng lên, chứng tỏ lực lượng lao động của tỉnh tăng lên.

4. Với độ tin cậy 95% ta kết luận điểm trung bình môn Thống kê của cả khóa nằm trong khoảng (7,26 ; 7,74) ($t_{149}(0,025) = 1,98$).

$EX = 7$ / $EX > 7$; $\alpha = 0,05$. Phương sai chưa biết, X phân phối chuẩn.

$t = 4,07 > 1,66 = t_{149}(0,05) \rightarrow$ Chấp nhận $EX > 7$ điểm.

5. Ở đây $DX = 9$, X phân phối chuẩn. Từ mẫu ta tính được $\bar{X} = 85,02$.

a) Chỉ số IQ trung bình là 85,02.

b) H: EX = 84 / K: EX > 84; $\alpha = 0,05$; $u(0,05) = 1,65$.

$u = \frac{85,02 - 84}{3}\sqrt{50} = 2,40 > 1,65 \rightarrow$ Chấp nhận EX > 84.

c) Với xác suất 90% chỉ số IQ trung bình thấp nhất là: 84,32.

d) $p = P(X \geq 84)$; $p^* = \frac{30}{50} = 0,6$; $p \in (0,44 ; 0,76)$.

6. a) 30%.

b) 37,56%.

7. Nếu số liệu là mẫu đại diện thì:

a) $p^* = 0,48$.

b) $p = 0,62$ / $p < 0,62$; $\alpha = 0,05$; $u = -11,2$

$\rightarrow$ Chấp nhận $p < 62\%$.

$p = 0,5$ / $p < 0,5$; $\alpha = 0,05 \rightarrow u = -1,549 > -1,65$

$\rightarrow$ Không kết luận $p < 50\%$ được.

c) p cao nhất là 50,13% (với xác suất 90%).

Nếu số liệu đã cho là số liệu tổng kết đầy đủ thì $p = 0,48$; do đó $p < 0,62$ và tất nhiên $p < 0,50$. Khi đó, việc ước lượng và kiểm định về p không cần đặt ra nữa.

8. a) $p^* = \frac{161}{644} = 0,25$. b) (0,222 ; 0,278).

c) $p = 0,40$ / $p < 0,40$; $\alpha = 0,05$; $u = -7,77 \rightarrow$ Chấp nhận tỷ lệ p đã giảm xuống.

d) Nếu số liệu là đầy đủ và chính xác thì không có việc ước lượng điểm, ước lượng khoảng, kiểm định giả thiết nữa. Ba câu hỏi trên được thay bởi câu: Hãy tính tỷ lệ tội phạm ở thanh thiếu niên sau khi có Nghị định 87/CP. Tỷ lệ này giảm

hay tăng so với 6 tháng đầu năm ? Khi đó p = 0,25 < 0,40 nên kết luận tỷ lệ p đã giảm.

9. Tính được $\overline{T}$ = 20,429 ; s = 9,8.

a) Thời gian tự học trung bình là 20,429 giờ ≈ 20 giờ 26 phút.

b) ET = 20 / ET > 20 ; α = 0,10 ; $t_{69}(0,10) = 1,29$

t = 0,3638 < 1,29 → Chưa thể kết luận ET > 20.

c) Thời gian tự học trung bình cao nhất là 22,788 giờ ≈ 22,79 giờ ≈ 22 giờ 47 phút. Khả năng đúng của kết luận trên là 95%, khả năng sai là 5%.

d) $p = P(T \geq 30) \rightarrow p^* = \dfrac{12}{70} = 0,1714 \approx 17\%$.

$p \in (0,0972 ; \ 0,2454)$; $p^*_{min} = 0,0972 = 9,7\%$.

10. Gọi p là xác suất khỏi bệnh = tỷ lệ người khỏi bệnh. $p^* = 0,733$.

p = 0,80 / p ≠ 0,8; α = 0,05; u(0,05) = 1,65; u(0,025) = 1,96

u = –2,14 → |–2,14| > 1,96 → Không chấp nhận p = 80%.

Do $p^* = 0,733 < 0,80$ nên ta chọn đối thiết p < 0,80.

p = 0,80 / p < 0,80; α = 0,05 → –2,14 < –1,65 → Chấp nhận p < 80%.

11. Ta tính được: $\overline{X}$ = 5,685 ; s = 0,8504 ; $\hat{s}$ = 0,8547.

a) EX = 5,5 / EX > 5,5 ; α = 0,05. DX chưa biết, không phân phối chuẩn, n = 100 đủ lớn. u = 2,165 → Chấp nhận EX > 5,5 → Nên mở siêu thị ở khu dân cư M.

b) Gọi $p = P(X \geq 5,5) \rightarrow p^* = \dfrac{70}{100} = 0,70$.

Với xác suất 95%; $p \in (0,61 ; 0,79)$

c) EX = EY / EX < EY; α = 0,05. DX, DY chưa biết. Không có giả thiết chuẩn, $n_1 = n_2 = 100$ đủ lớn.

$u = -0,823 > -1,65 \rightarrow$ Chấp nhận EX = EY.

12. Gọi p_{bd}, p_{kl} tương ứng là tỷ lệ cây bồ đề, keo lai sống.

$p_{bd} = p_{kl}$ / $p_{bd} > p_{kl}$; $\alpha = 0,10$; $u(0,10) = 1,28$.

$u = 3,55 \rightarrow$ Chấp nhận $p_{bd} > p_{kl}$, cây bồ đề có tỷ lệ sống cao hơn.

13. Gọi p_{cl}, p_{dr} tương ứng là tỷ lệ người thích dùng dầu gội Clear và Dr.men.

$p_{cl} = p_{dr}$ / $p_{cl} < p_{dr}$; $\alpha = 0,02$; $u(0,02) = 2,05$.

$u = -0,824 > -2,05 \rightarrow$ Chấp nhận $p_{cl} = p_{dr}$, tức là chưa đủ cơ sở để khẳng định người tiêu dùng không thích dùng dầu gội Clear bằng Dr. Men.

14. EX = EY / EX < EY ; $\alpha = 0,10$; $u(0,10) = 1,28$;

DX = DY = 9

$u = -5,33 \rightarrow$ Chấp nhận EX < EY.

15. Tính được $\bar{X} = 6,13$ $s_x = 1,695$ $\hat{s}_x = 1,704$

$\bar{Y} = 6,75$ $s_y = 1,682$ $\hat{s}_y = 1,690$

a) Vì $\bar{X} < \bar{Y}$ nên ta chọn đối thiết EX < EY:

EX = EY / EX < EY; $\alpha = 0,05$. DX, DY chưa biết nhưng DX = DY; X, Y tuân theo luật chuẩn; $t_{198}(0,05) = 1,65$.

$t = -2,583 < -1,65 \rightarrow$ Chấp nhận EX < EY

b) Gọi p_A, p_B là tỷ lệ điểm khá giỏi ở hai trường A, B tương ứng.

$p_A^* = 0,39$; $p_B^* = 0,58$; $p_A^* < p_B^*$ nên ta chọn $p_A < p_B$

$p_A = p_B$ / $p_A < p_B$ $\alpha = 0,05$; $u(0,05) = 1,65$.

$u = -2,676 < -1,65 \rightarrow$ Chấp nhận $p_A < p_B$.

c) So sánh điểm trung bình $p_A^* = 0,45$; $p_B^* = 0,32$.

$p_A = p_B$ / $p_A > p_B$; $\alpha = 0,05$

$u = 1,884 > 1,65 \rightarrow$ Chấp nhận $p_A > p_B$.

d) So sánh điểm giỏi.

Ước lượng điểm là: $p_A^* = 0,21$; $p_B^* = 0,30$.

Rõ ràng hai ước lượng điểm là khác nhau. Nếu $p_A = p_B$ thì sự khác nhau này là ngẫu nhiên, còn nếu $p_A \neq p_B$ thì sự khác nhau của hai ước lượng điểm thuộc về bản chất:

$p_A = p_B$ / $p_A \neq p_B$; $\alpha = 0,05$; $u(0,025) = 1,96$.

$u = -1,452 \rightarrow |-1,452| < 1,96 \rightarrow$ Chấp nhận $p_A = p_B$.

Sự khác nhau giữa hai ước lượng điểm (21% và 30%) là sự khác nhau ngẫu nhiên. Chưa đủ cơ sở để nói tỷ lệ điểm giỏi thực của hai trường là khác nhau. Có lẽ ta cũng có kết luận tương tự khi so sánh điểm kém giữa hai trường.

e) Như vậy, với mức ý nghĩa 5% ta có thể nói chất lượng dạy và học môn tiếng Anh của trường B tốt hơn trường A, thể hiện ở chỗ: Điểm trung bình trường B cao hơn, tỷ lệ điểm khá giỏi của trường B cũng cao hơn, tỷ lệ điểm trung bình trường B thấp hơn.

16. $n_1 = 20$; $n_2 = 15$. Ta chỉ có thể dùng tiêu chuẩn Mann–Whitney.

H: EX = EY / K: EX ≠ EY; $\alpha = 0,10$; $u(0,05) = 1,65$.

$R_2 = 131$; $R_1 = 499$; $U = 11$; $u = -4,67 \rightarrow$ Chấp nhận EX ≠ EY.

Phần lớn mẫu I đứng cuối, cho nên dễ thấy $\bar{X} > \bar{Y}$, do đó chấp nhận EX > EY, tức là doanh số trung bình của nhóm I cao hơn nhóm II.

17. Đây là số liệu được cho theo từng cặp, nên chỉ có thể dùng tiêu chuẩn Wilcoxon.

$EX = EY$ / $EX \neq EY$; $\alpha = 0,05$; $u(0,025) = 1,96$

$n^+ = 18$; $T^- = 62,5$; $T^+ = 108,5$; $T = 62,5$; $u = -1,0$

→ Chấp nhận $EX = EY$. Ta bác bỏ ý kiến nêu ra.

Với kết luận trên ta thấy người dân thuộc nhóm khảo sát ở đây đã có ý thức tiết kiệm điện tốt. Lượng điện mà họ tiêu thụ là lượng cần thiết để phục vụ cuộc sống của họ.

18. Số liệu đã cho thuộc dạng phụ thuộc theo từng cặp, nên phải dùng tiêu chuẩn Wilcoxon.

$EX = EY$ / $EX \neq EY$; $\alpha = 0,02$; $u(0,01) = 2,33$.

$n^+ = 17$; $T^- = 8$; $(T^+ = 145)$; $T = 8$; $u = -3,24$

→ Chấp nhận $EX \neq EY$.

Do phần lớn $d_i > 0$, chỉ có 3 giá trị $d_i < 0$ nên dễ thấy $\overline{X} > \overline{Y}$, vì thế ta chấp nhận $EX > EY$, tức là học sinh lớp chuyên này có sự học lệch về các môn tự nhiên hơn.

19. Mô hình so sánh 2 giá trị trung bình.

Dùng tiêu chuẩn Mann–Whitney.

H: Hai mẫu thuần nhất; K: Hai mẫu không thuần nhất; $u(0,05) = 1,65$

$R_1 = 348$; $R_2 = 213$; $U = U_1 = 93$; $u = -1,52$

$|-1,52| < 1,65$ → Chấp nhận hai mẫu thuần nhất, nghĩa là về trung bình kết quả của nữ và nam là như nhau. Chưa có cơ sở để khẳng định kết quả của nữ tốt hơn nam.

20.

a) H: Số liệu mẫu phù hợp với 4 tỷ lệ đã cho ~ Chấp nhận báo cáo của trường.

K: Số liệu mẫu không phù hợp với 4 tỷ lệ đã cho ~ Không chấp nhận báo cáo của trường.

Vì $m_1 = 2 < 5$, nên ta gộp tỷ lệ giỏi và khá lại với nhau thành 35%, và $m_1 + m_2 = 2 + 10 = 12$.

Do đó: $k = 3$, $\chi_2^2(0,05) = 5,99$.

$\chi^2 = 56,86 > 5,99$. Theo (II.27), chúng ta không chấp nhận báo cáo của nhà trường.

b) Kiểm định tỷ lệ điểm trung bình:

$p = 0,50 / p \neq 0,50$; $\alpha = 0,05$

$u = -0,4$ nên ta chấp nhận tỷ lệ điểm trung bình 50% theo báo cáo.

Kiểm định tỷ lệ điểm kém: $p = 0,15 / p < 0,15$; $\alpha = 0,05$

$u = 7,0 \rightarrow$ Không chấp nhận tỷ lệ điểm kém 15% theo báo cáo.

c) Với xác suất 95% tỷ lệ điểm kém thấp nhất cũng là 30,4%.

21. Đây là mô hình so sánh ba tỷ lệ: $\chi_2^2(0,05) = 5,99$

$\chi^2 = 5,635 \rightarrow$ Chấp nhận ba tỷ lệ như nhau.

22.

a) H: Không có ảnh hưởng gì ~ Hai kết quả độc lập.

K: Có ảnh hưởng ~ Hai kết quả có sự phụ thuộc.

$\chi^2 = 142.0,3797 = 53,917 > 7,78 = \chi_4^2(0,10)$.

Kết quả các môn tự nhiên có ảnh hưởng tới kết quả môn ngoại ngữ.

b) Ta thấy kết luận trên phù hợp với thực tiễn. Phần lớn những người học tự nhiên tốt thì học ngoại ngữ cũng không

tồi (trừ phần thuộc về năng khiếu). Bởi lẽ, học tự nhiên tốt thì họ có tư duy logic tốt, do đó tiếp thu tư duy của ngôn ngữ sẽ nhanh. Điều đó rất giúp ích cho học ngoại ngữ.

23. Đây là mô hình kiểm tra độc lập.

H: Kết quả thi độc lập với khoa.

$\chi^2 = 7,32 < 12,6 = \chi^2_6(0,05) \rightarrow$ Chấp nhận kết quả thi độc lập với khoa (miền (II.28)).

24. H: Tình trạng hôn nhân của vợ và của chồng độc lập với nhau.

$\chi^2 = 542.0,1476 = 79,9992 > 9,49 = \chi^2_4(0,05)$

Vậy, tình trạng hôn nhân của vợ và của chồng trước ngày cưới có sự phụ thuộc nhau.

25. Theo bảng số liệu đã cho ta có r = 2 (ốm đau và không ốm đau).

Bài toán phù hợp cho cả 2 mô hình: Kiểm tra độc lập hoặc so sánh ba tỷ lệ ($\chi^2_2(0,05) = 5,99$)

$\chi^2 = 10,3 > 5,99 \rightarrow$ Nghỉ ốm có phụ thuộc vào tuổi tác.

26.

a) H: Tình trạng sức khỏe của trẻ ở tỉnh B cũng như ở tỉnh A.

K: Tình trạng sức khỏe của trẻ ở tỉnh B khác với ở tỉnh A.

(Hoặc tương đương H: Số liệu ở tỉnh B phù hợp với các tỷ lệ đã cho ở tỉnh A).

$k = 3$; $p_1 = 0,11$; $p_2 = 0,33$; $p_3 = 1 - (0,11 + 0,33) = 0,56$

$n = 3000$; $m_1 = 270$; $m_2 = 830$; $m_3 = 1900$;

$\chi^2 = 65,58 > 4,64 = \chi^2_2(0,10)$

→ Tình trạng sức khỏe của trẻ ở tỉnh B khác ở tỉnh A.

b) Gọi p là tỷ lệ trẻ bị suy dinh dưỡng nặng ở tỉnh B.

$p = 0,11 / p \neq 0,11; u(0,05) = 1,65$

$u = -3,5; |-3,5| > 1,65 \rightarrow p \neq 11\%$

Vì $p^* = 0,09 < 0,11$ nên ta chọn đối thiết $p < 0,11$.

$p = 0,11 / p < 0,11; u(0,10) = 1,28$

$u = -3,5 < -1,28 \rightarrow$ Chấp nhận $p < 0,11$. Ở tỉnh B tỷ lệ trẻ suy dinh dưỡng nặng thấp hơn ở tỉnh A.

27.

a) Ta thấy X tăng thì Y cũng tăng, X giảm thì Y cũng giảm.

Tức là tỷ lệ gia tăng số người chết thay đổi cùng chiều hướng với tỷ lệ gia tăng xe máy.

b) $\bar{X} = 21,34$; $s_x = 13,03$; $\bar{Y} = 21,48$; $s_y = 16,44$;

$\overline{XY} = 663,86 \rightarrow r = 0,96$.

Với $r = 0,96$ chứng tỏ Y phụ thuộc tuyến tính rất chặt vào X và sự phụ thuộc là đồng biến. Điều đó càng chứng minh cho nhận xét nêu trên.

28.

a) $r = 0,967$

b) Sự phụ thuộc giữa số người nhiễm HIV và số người tiêm chích ma tuý là rất chặt và sự phụ thuộc là đồng biến.

c) $y = 0,519x + 14,26$

d) Đường hồi quy nhận được có thể dùng để dự doán số người nhiễm HIV khi cho biết số người tiêm chích ma túy ở

địa phương đó. Vì r = 0,967, khá gần 1, chứng tỏ mức độ phụ thuộc tuyến tính giữa y và x là rất chặt, dự đoán sẽ rất tốt.

Chẳng hạn, nếu có 265 người tiêm chích ma túy, thì ta dự đoán ở địa phương đó sẽ có 152 người nhiễm HIV.

e) $s^2_{y/x} = 102,96$.

29. Chọn mỗi khoảng giá trị của X một điểm đại diện (điểm giữa của mỗi khoảng).

Ở đây: $m_i = 1$, $\forall i$; $\bar{X} = 1,1$; $s_x = 0,513$; $\bar{Y} = 35$;

$s_y = 21,61$; $\overline{XY} = 49,5$

a) r = 0,992.

b) Tỷ lệ người tham gia du lịch phụ thuộc tuyến tính rất chặt vào mức thu nhập thêm của họ.

c) y = 41,788x – 10,967.

30.

a) H: Kết quả đánh giá độc lập với khoa được hỏi; $\chi^2_4(0,05) = 9,49$.

$\chi^2 = 7,45 < 9,49 \rightarrow$ Chấp nhận tính độc lập.

b) $p_A = 0,55$ / $p_A > 0,55$; $\alpha = 0,05$; $u(0,05) = 1,65$

$u = 0,71 \rightarrow$ Chấp nhận $p_A = 0,55$.

Như vậy, mặc dù ước lượng điểm $p^*_A = 60\%$, nhưng chưa đủ cơ sở để khẳng định $p_A > 55\%$.

c) Tỷ lệ cao nhất là: 16,65% $\quad (p^* = \dfrac{16}{133} = 0,12)$.

d) So sánh 3 tỷ lệ.

Số liệu mỗi cột (ứng với mỗi khoa) ta chia làm hai phần: đánh giá tốt và không đánh giá tốt (gồm đánh giá trung bình, yếu hoặc không quan tâm).

$\chi^2 = 0,4877 < 5,99 = \chi^2_2(0,05)$. Ba tỷ lệ như nhau.

31. Tính được $\bar{X} = 3,5$; $s_x^2 = 1,85$; $\bar{Y} = 69$; $s_y^2 = 195$.

a) $r = 0,842$.

b) $y = 8,6446x + 39,256$.

c) Đường hồi quy trên có thể dùng để dự đoán số thu nhập thêm khi biết số trẻ em trên 7 tuổi nhưng dưới 16 tuổi. Vì $r = 0,842$, khá gần 1, nên mối tương quan giữa hai tiêu thức khá chặt. Dự báo sẽ khá tốt.

d) $s^2_{y/x} = 56,745$.

PHỤ LỤC I

MỘT SỐ KHÁI NIỆM CỦA XÁC SUẤT

Do hạn chế về thời gian (30 tiết cho môn học này) nên chúng ta không thể tìm hiểu sâu các khái niệm và kết quả cơ bản của lý thuyết xác suất, nhưng chúng ta cũng không thể không nhắc gì đến chúng, bởi vì chúng được dùng trong phần Thống kê. Vì vậy, phần này sẽ giới thiệu một số khái niệm cần thiết, giải thích ý nghĩa thực tiễn, ý nghĩa đời thường của chúng giúp bạn đọc hiểu vấn đề hơn. Phần này không đặt ra yêu cầu thực hành. Chẳng hạn, khi nói đến xác suất, kỳ vọng, phương sai,... bạn đọc hiểu được chúng là gì, khi viết $P(A) = 0,90$; $P(X > a) = 95\%$; $EX = 45$; $\sigma = 3$... yêu cầu bạn đọc phát biểu được bằng lời các đẳng thức trên và hiểu được ý nghĩa thực tiễn của các mệnh đề đó. Không yêu cầu bạn đọc phải biết tính toán, giải thích tại sao chúng ta có các kết quả đó.

1. Phép thử và biến cố

Ta gieo một đồng tiền trên mặt phẳng, hoặc trọng tài cho 2 đội chọn sân và chọn bóng trong một trận bóng đá...; gieo một con xúc xắc trên mặt phẳng; bắn một viên đạn vào bia để tính điểm; đo nhiệt độ cho bệnh nhân; chọn ngẫu nhiên một

sản phẩm; phỏng vấn ngẫu nhiên một cử tri, một khách hàng; v.v... đều là các phép thử.

Phép thử mà ta không thể khẳng định được trước kết quả của nó, được gọi là phép thử ngẫu nhiên.

Các tình huống có thể xảy ra khi phép thử được thực hiện được gọi là các biến cố.

Các tình huống có thể xảy ra của phép thử, nhưng không thể phân tách nhỏ hơn được nữa gọi là các biến cố sơ cấp.

Ví dụ như phép thử gieo con xúc xắc trên mặt phẳng: "Xuất hiện mặt k chấm", k = 1, 2,..., 6 là 6 biến cố sơ cấp. "Xuất hiện mặt có số chấm chẵn" là biến cố (nhưng không phải biến cố sơ cấp). Phép thử bắn một viên đạn vào bia để tính điểm: "Bắn được k điểm"; k = 0,1,...,10 là 11 biến cố sơ cấp; nhưng "Bắn đạt yêu cầu" ; "Bắn đạt điểm giỏi",... cũng là các biến cố (song không phải biến cố sơ cấp).

Người ta phân chia biến cố thành ba loại sau:

– Biến cố không thể (ký hiệu ϕ) là biến cố không thể xảy ra khi phép thử được thực hiện.

– Biến cố chắc chắn (ký hiệu Ω) là biến cố nhất định xảy ra khi phép thử được thực hiện.

– Biến cố ngẫu nhiên (ký hiệu A, B, C,...) là biến cố có thể xảy ra và cũng có thể không xảy ra khi phép thử được thực hiện.

Trong ba loại biến cố trên, biến cố ngẫu nhiên là phổ biến hơn cả. Nghiên cứu các biến cố ngẫu nhiên chính là đối tượng nghiên cứu đầu tiên của Lý thuyết Xác suất và Thống kê Toán học.

2. Định nghĩa xác suất

Như vậy, biến cố ngẫu nhiên A có hai khả năng (xảy ra và không xảy ra), biết được khả năng này ta suy ra ngay khả năng kia. Vì vậy, để nghiên cứu biến cố ngẫu nhiên A chúng ta xác định (hoặc đặc trưng) tương ứng với biến cố A một con số, ký hiệu P(A), nói lên khả năng xảy ra biến cố A (do đó 1 – P(A) là con số nói lên khả năng không xảy ra biến cố A) và được gọi là xác suất của biến cố A. (P là viết tắt của từ Probability, đôi khi người ta viết Prob(A)).

Như vậy, xác suất của biến cố A chỉ là một con số, nói lên khả năng xảy ra hiện tượng A đang xét. Mà khả năng xảy ra lại là một khái niệm rất đời thường, người ít chữ vẫn thường dùng, bản thân mỗi chúng ta cũng đã từng dùng nó. Do đó, xác suất là thuật ngữ mới, nhưng thực chất lại là khái niệm rất quen thuộc: khả năng xảy ra.

Rõ ràng, người ta dùng con số từ 0 đến 100% để chỉ khả năng xảy ra, cho nên chúng ta thấy ngay các tính chất sau của xác suất:

$0 \le P(A) \le 1 = 100\%$

$P(\phi) = 0$; $P(\Omega) = 1$; $0 < P(A) < 1$.

P(A) càng gần 1 thì khả năng xảy ra biến cố A càng nhiều, càng lớn.

P(A) càng gần 0 thì khả năng xảy ra biến cố A càng ít, càng nhỏ.

Khi viết P(A) = 0,95 ta hiểu ngay rằng: khả năng xảy ra biến cố A là 95%, còn khi viết P(B) = 1% ta hiểu rằng khả năng xảy ra biến cố B chỉ là 1%. Tuy A và B đều là biến cố ngẫu nhiên, nhưng chúng ta trông chờ (mong chờ) A xảy ra, chứ không hy vọng B xảy ra.

Rõ ràng: $P(A) + P(\bar{A}) = 1$

(Khả năng xảy ra A cộng với khả năng không xảy ra A là 100%).

Một câu hỏi đặt ra là: Tìm con số P(A) như thế nào?

Một cách tính P(A) là dùng định nghĩa cổ điển, nghĩa là:

$$P(A) = \frac{\text{Số biến cố sơ cấp thuận lợi cho A}}{\text{Số biến cố sơ cấp có thể xảy ra khi phép thử được thực hiện}}$$

Trong đó, biến cố sơ cấp thuận lợi cho A, nghĩa là nếu biến cố sơ cấp đó xảy ra thì suy ra A xảy ra.

Ví dụ 1

a) Trong một lớp có 100 sinh viên, trong đó có 65 nữ. Chọn ngẫu nhiên một sinh viên trong lớp. Khi đó:

$$P(\text{Chọn được nữ sinh viên}) = \frac{C_{65}^1}{C_{100}^1} = \frac{65}{100}$$

Nhưng phân số $\frac{65}{100}$ này lại chính là tỷ lệ nữ trong lớp.

b) Trong một trường THPT có tỷ lệ học sinh bị cận thị là 30%. Chọn ngẫu nhiên một học sinh trong trường. Khả năng chọn được học sinh bị cận thị sẽ là 30%.

Như vậy, trong nhiều tình huống xác suất lại chính là tỷ lệ, nghĩa là P(A) chính là tỷ lệ dấu hiệu A trong tập đang xét.

(Đưa vào định nghĩa cổ điển trên, tác giả chỉ muốn dẫn dắt bạn đọc đến một khái niệm dễ hiểu khác của xác suất là tỷ lệ, chứ không yêu cầu bạn đọc biết dùng định nghĩa này để tính xác suất, do đó tác giả không trình bày kỹ định nghĩa này).

Tóm lại: *Xác suất của biến cố A chính là khả năng xảy ra A hoặc là tỷ lệ các phần tử dấu hiệu A trong tập đang xét.*

3. Biến ngẫu nhiên

Một biến (hay một đại lượng) nhận các giá trị của nó với xác suất tương ứng nào đó, được gọi là biến ngẫu nhiên (đại lượng ngẫu nhiên). Ta ký hiệu biến ngẫu nhiên là: X, Y, Z,...

Nói đến biến hay đại lượng chắc chắn bạn đọc không xa lạ gì. Nhưng điểm mới ở đây là gì ? Biến mà chúng ta thường dùng ở phổ thông là biến tất định. Một giá trị nào đó đối với biến tất định chỉ có thể là nhận hoặc không nhận, hay nói cách khác biến tất định nhận giá trị của nó với xác suất 1 hoặc 0. (Thông thường, ta chỉ nói đến giá trị biến nhận, không nói tới giá trị biến không nhận). Còn biến ngẫu nhiên nhận giá trị của nó với xác suất nào đó lớn hơn 0 và nhỏ hơn 1.

Ví dụ 2

a) Sau khi làm bài thi xong, sinh viên A cảm thấy bài làm khá hoàn chỉnh, chỉ mắc một vài sai sót nhỏ. A nói với các bạn trong phòng là: "Mình được 9 điểm". A nói như vậy được hay không?

Nếu hiểu theo cách nói tất định, nghĩa là A khẳng định chắc chắn mình được 9 điểm, hay mình được 9 điểm với xác suất 1 thì rõ ràng không ổn, không chặt chẽ, không chính xác. Người ta có thể đặt câu hỏi: Đáp án và thang điểm A chưa biết, làm sao A có thể chắc chắn như vậy? Phải chăng A "móc ngoặc" với người chấm bài? Vì thế, không thể hiểu theo nghĩa tất định, mà chỉ có thể hiểu theo nghĩa ngẫu nhiên, tức là A đạt điểm 9 với khả năng cao nhất trong các khả năng

có thể. Khi đó, nếu gọi X là điểm thi của sinh viên A. X là biến ngẫu nhiên, X có thể được mô tả như sau:

Giá trị của X	0	1	2	...	6	7	8	9	10
Xác suất tương ứng	0	0	0	...	0	0,05	0,10	0,80	0,05

Nhìn vào bảng trên, ta thấy chắc chắn A đạt từ điểm khá trở lên, và khả năng A được 9 điểm là cao nhất và bằng 80%, chắc chắn A không bị điểm dưới 7. Khả năng A đạt 8 điểm, 10 điểm cũng có, nhưng ít.

Nhưng sau khi đã có kết quả thông báo trên bảng của khoa (hoặc của Phòng Đào tạo). A thấy mình được điểm 9. Bây giờ A có thể nói với các bạn trong phòng: "Mình được 9 điểm", tức là nói theo nghĩa tất định, chứ không phải theo nghĩa ngẫu nhiên nữa. Bây giờ nếu A nói "Mình được 9 điểm với xác suất 80%" thì lại là buồn cười và không đúng.

b) Gọi Y là biến ngẫu nhiên chỉ số mặt sấp xuất hiện khi ta gieo 2 đồng tiền cân đối, đồng chất trên mặt phẳng. Rõ ràng không thể mô tả Y theo nghĩa tất định. Ta có thể mô tả Y như sau:

Y nhận các giá trị	0	1	2
Xác suất tương ứng	0,25	0,50	0,25

Nhìn vào bảng, ta thấy P(Y = 1) là cao nhất (50%). Vậy trước khi gieo, nếu đoán, ta nên đoán sẽ có 1 mặt sấp, 1 mặt ngửa. Đoán như vậy khả năng đúng sẽ cao nhất.

(Ở đây bạn đọc phải công nhận 3 giá trị xác suất được cho ở bảng trên, hoặc bạn đọc có thể tìm chúng bằng định nghĩa cổ điển).

Trong truyện Trạng Quỳnh, có một câu chuyện như sau: Đền thờ bà Chúa Liễu Hạnh có tiếng là rất thiêng. Phật tử bốn phương đến cúng lễ và bỏ vào Đền khá nhiều tiền. Trạng Quỳnh dóng vai một sỹ tử nghèo, đang thiếu tiền lộ phí trên đường đi thi. Khi qua Đền, Trạng Quỳnh vào Đền khấn lễ. Trước khi gieo Trạng Quỳnh đã khấn đại ý rằng: Nếu Chúa đồng ý cho vay một nửa thì hãy cho nhất âm nhất dương, còn nếu rơi vào trường hợp khác thì xin cho vay một phần tư.

Qua câu chuyện này cho thấy Trạng Quỳnh đã biết tính khả năng xảy ra khá tốt. Ông đã chọn tình huống xảy ra cao nhất để lấy nửa số tiền, và nếu không may mắn thì cũng được một phần tư số tiền.

Qua hai ví dụ trên bạn đọc có thể thấy rằng việc đưa ra khái niệm biến ngẫu nhiên là rất cần thiết.

4. Phân phối xác suất

Để xác định biến ngẫu nhiên, rõ ràng ta phải chỉ rõ hai thông tin: Giá trị mà biến ngẫu nhiên nhận và xác suất nhận giá trị đó tương ứng. Một bảng với hai thông tin như vậy được gọi là bảng phân phối xác suất, thường được gọi gọn là phân phối xác suất. Ta có:

X	x_1	x_2	. . . . .	x_n
$P(X = x_i)$	p_1	p_2	. . . . .	p_n

(1)

$$\sum_{i=1}^{n} p_i = 1$$

(Nếu trình bày kỹ hơn, như trong chương I thì để xác định biến ngẫu nhiên X ta đã dùng 3 khái niệm: bảng phân phối, hàm mật độ, hàm phân phối. Ở đây, để đơn giản, tác giả dùng một thuật ngữ chung là phân phối xác suất.

Giả sử ta có một phân phối xác suất nào đó, ký hiệu là N. Khi ta nói: "biến ngẫu nhiên N", hoặc "phân phối N", hoặc "biến ngẫu nhiên có phân phối N", thì ba cách nói trên sẽ được hiểu như nhau.

5. Kỳ vọng

Kỳ vọng của biến ngẫu nhiên X là một con số, ký hiệu EX (E là viết tắt của từ Expectation); chẳng hạn $EX = \sum_i x_i p_i$, nếu X có phân phối dạng (1).

Ý nghĩa:

Kỳ vọng của biến ngẫu nhiên X chính là giá trị trung bình mà biến ngẫu nhiên X nhận.

(Theo bạn, kỳ vọng EX là con số như thế nào? Có phải từ 0 đến 1 như xác suất không?)

Tính chất đơn giản của kỳ vọng:

EC = C (C là hằng số), E(CX) = CEX;

$E(X \pm Y) = EX \pm EY$, $E(aX + b) = aEX + b$.

Bạn đọc có thể lý giải được các tính chất này nếu bạn dùng các tính chất khi tính giá trị trung bình.

Như vậy, kỳ vọng thực ra lại là một khái niệm rất quen thuộc. Bạn đã dùng kỳ vọng chưa? Và dùng từ khi nào? Hãy chỉ ra các tình huống trong thực tế có dùng đến kỳ vọng.

Bạn đọc hãy trả lời các câu hỏi vừa đặt ra trước khi đọc tiếp phần sau.

Thực tế chúng ta đã dùng kỳ vọng từ khi học lớp 1. Bởi lẽ ở lớp 1 thầy cô đã dùng điểm trung bình để xếp thứ về kết quả học tập cho học sinh. Mà điểm trung bình cộng (trung bình số học) là một trường hợp riêng của kỳ vọng toán (nếu

$p_i = \frac{1}{n}$, $\forall i = \overline{1, n}$ thì $\sum_{i=1}^{n} x_i p_i = \frac{1}{n}\sum_{i=1}^{n} x_i$). (Tất nhiên ở lớp 1 chúng ta chưa thể tự tính được, mà thầy cô đã tính cho). Và giá trị kỳ vọng đó theo chúng ta trong suốt cả quãng đời đi học, từ phổ thông, cao đẳng, đại học, cao học.

Câu hỏi đặt ra là: Vì sao ở bậc Tiểu học người ta dùng trung bình số học, còn từ bậc THCS trở đi người ta dùng trung bình có hệ số (trung bình có trọng lượng)? Để lý giải, ta lấy tình huống đơn giản, cụ thể sau của hai học sinh ở bậc THPT:

Điểm / Học sinh	Kiểm tra 15'	Kiểm tra miệng	Kiểm tra 1 tiết	Thi học kỳ
Học sinh A	10			2
Học sinh B		2		10

Tính theo trung bình số học, ta có:

Học sinh A: $\frac{10+2}{2} = 6$

Học sinh B: $\frac{2+10}{2} = 6$

Nếu ghi vào học bạ cuối kỳ môn học đó, cả hai học sinh A và B đều điểm 6, rõ ràng là không ổn. Sở dĩ xảy ra tình trạng như vậy là vì khi tính trung bình cộng ta đã đánh đồng vai trò các điểm như nhau. Nhưng thực ra điểm 10 của học sinh B trong bài thi học kỳ khác xa về sự hiểu biết, về khối lượng kiến thức so với điểm 10 của học sinh A trong một bài kiểm tra 15' (hoặc kiểm tra miệng).

Nếu tính theo trung bình có hệ số, ta có:

Học sinh A: $10.\frac{1}{4} + 2.\frac{3}{4} = 4$

Học sinh B: $2.\frac{1}{4} + 10.\frac{3}{4} = 8$

(Thực chất là $x_1.p_1 + x_2.p_2$, mà $p_1 = \frac{1}{4}$; $p_2 = \frac{3}{4}$; $p_1 + p_2 = 1$)

Ghi vào học bạ: học sinh A điểm 4, học sinh B điểm 8. Kết quả đó phản ánh chân thực, khách quan và đúng đắn trình độ của học sinh hơn nhiều.

Ở bậc cao đẳng, đại học, cao học người ta vẫn tính điểm trung bình của từng học kỳ, của cả khóa học để xét thi đua, khen thưởng, xét học bổng, xếp loại tốt nghiệp. Điểm trung bình được tính theo công thức:

$$\text{Điểm TB} = \sum_i x_i \left(\frac{w_i}{\sum_i w_i} \right)$$

trong đó: x_i – điểm thi; w_i – số đơn vị học trình của môn thứ i.

Công thức trên cũng chỉ là trường hợp riêng của công thức kỳ vọng với $P_i = \frac{w_i}{\sum w_i}$

Như vậy, giá trị kỳ vọng là một đặc trưng hữu hiệu được dùng trong các trường học.

Ngoài trường học, trên thực tế có rất nhiều tình huống có dùng đến kỳ vọng. Ví dụ: Mật độ dân số; thu nhập GDP bình quân tính theo đầu người, thu nhập trung bình của những người lao động trong một ngành nghề nào đó; năng suất của

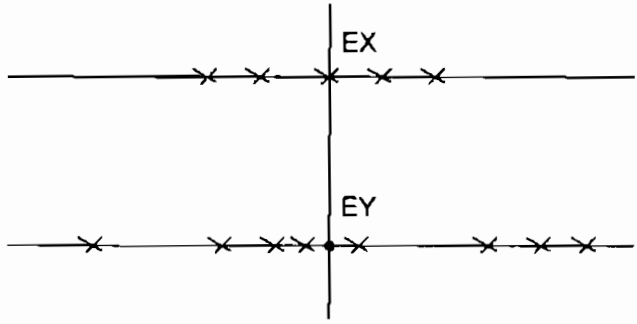
một huyện, một tỉnh; nhiệt độ, độ ẩm trung bình của vùng nào đó trong ngày mai, ngày kia, v.v...

Nói chung, không bao giờ ta dùng một giá trị ngẫu nhiên mà biến nhận để nói về biến ngẫu nhiên đó, mà người ta thường dùng giá trị trung bình (dùng kỳ vọng). Có thể nói, kỳ vọng là giá trị phản ánh trung thực và khách quan nhất trong tất cả các giá trị trung bình cộng.

6. Phương sai

Là biến ngẫu nhiên, giá trị của nó luôn thay đổi, lúc thế này, lúc thế khác (nhưng có quy luật). Vậy, ngoài giá trị trung bình – một đặc trưng hữu hiệu, chúng ta cần chỉ ra mức độ thay đổi của biến ngẫu nhiên.

Giả sử ta có 2 biến ngẫu nhiên X và Y. Các giá trị của chúng được mô tả bởi các điểm "x" trên 2 trục số thực như sau:



Trong tình huống trên: EX = EY, tức là 2 biến ngẫu nhiên X và Y có cùng kỳ vọng, nhưng bức tranh về các giá trị của X và các giá trị mà Y nhận thì khác hẳn nhau. Các giá trị của X tập trung quanh EX, nhưng các giá trị của Y thì phân tán, tản

mát so với kỳ vọng EY của nó. Chúng ta cần xây dựng một đặc trưng khác phản ánh tình huống trên. Đó chính là phương sai, được ký hiệu là DX và được xác định như sau:

$$DX = E(X - EX)^2,$$

trong đó:

$(X - EX)$: khoảng cách từ giá trị của biến X tới giá trị kỳ vọng EX (sẽ có giá trị dương, giá trị âm).

$(X - EX)^2$: bình phương khoảng cách từ X tới EX (để mất dấu dương, âm).

$E(X - EX)^2$: trung bình của các bình phương khoảng cách.

Ý nghĩa:

Phương sai DX của biến ngẫu nhiên X là một số không âm (không có đơn vị), nói lên mức độ tập trung (hoặc mức độ tản mát, phân tán) của các giá trị của biến ngẫu nhiên X so với giá trị trung bình của nó. Nếu DX càng nhỏ, các giá trị của biến ngẫu nhiên X càng tập trung quanh EX, khi đó sự chênh lệch giữa các giá trị mà biến ngẫu nhiên X nhận sẽ nhỏ; còn nếu DX càng lớn thì các giá trị của biến ngẫu nhiên X càng phân tán, khi đó sự sai khác giữa giá trị lớn và giá trị nhỏ của biến ngẫu nhiên X sẽ khá lớn. Vì thế, thay vì nói "Phương sai của biến ngẫu nhiên X" ta nói "Bình phương độ tản mát của biến ngẫu nhiên X".

Tính chất đơn giản của phương sai:

$0 \le DX = E(X - EX)^2 = EX^2 - (EX)^2$

(Thật vậy: $E(X - EX)^2 = E(X^2 - 2X.EX + (EX)^2)$

$= EX^2 - 2EX.\ EX + E(EX)^2$

$= EX^2 - 2(EX)^2 + (EX)^2$

$= EX^2 - (EX)^2$).

trong đó: nếu phân phối của X được cho dưới dạng (1) thì $EX^2 = \sum_i x_i^2 p_i$.

$DC = 0$

$DCX = C^2DX$

$D(aX + b) = a^2DX$

Dễ dàng chứng minh các tính chất này bằng định nghĩa và tính chất của kỳ vọng, chẳng hạn:

$$D(aX + b) = E\{(aX + b) - E(aX + b)\}^2$$
$$= E\{aX + b - aEX - b\}^2$$
$$= E\{a(X - EX)\}^2 = Ea^2(X - EX)^2$$
$$= a^2 E(X - EX)^2 = a^2 DX$$

Ngoài cách ký hiệu EX, DX, để đơn giản người ta còn ký hiệu dưới dạng tham số. Thay cho EX ta viết là μ, thay cho DX ta viết là σ^2.

Ta gọi $\sigma = \sqrt{DX}$ là độ lệch tiêu chuẩn của biến ngẫu nhiên X.

(Độ lệch tiêu chuẩn có cùng đơn vị như biến ngẫu nhiên X hoặc như đơn vị của EX).

Phần này không yêu cầu thực hành tính EX, DX. Bạn đọc có thể tự tính thử đối với ví dụ 2. (Kết quả: $EX = 8,85$; $EX^2 = 78,3225$; $DX = 0,3275$; $EY = 1$; $EY^2 = 1,5$; $DY = 0,5$).

7. Mode

Nếu phân phối của biến ngẫu nhiên X được cho dưới dạng (1) thì Mode của X là một giá trị mà tại đó biến ngẫu nhiên nhận với xác suất lớn nhất, ta ký hiệu là ModX.

Trong ví dụ 2: ModX = 9; Mod Y = 1.

8. Một vài phân phối cần dùng

– Phân phối chuẩn, ký hiệu $N(\mu; \sigma^2)$ (N là từ viết tắt của từ normal), nhận giá trị $(\infty < x < + \infty)$. Ta nói "X là biến ngẫu nhiên chuẩn" hoặc "X có phân phối chuẩn" hoặc "Biến ngẫu nhiên X tuân theo luật chuẩn" đều được hiểu như nhau, và được ký hiệu $X = N(\mu; \sigma^2)$ trong đó: tham số thứ nhất μ được hiểu là EX, tham số thứ hai σ^2 được hiểu là DX.

Ký hiệu $u(\alpha)$ là một giá trị xác định từ đẳng thức:

$$P\left(\frac{X-\mu}{\sigma} \geq u(\alpha)\right) = \alpha$$

$$P\ (X \geq \mu + \sigma.\ u(\alpha)) = \alpha \qquad (2)$$

Hoặc tương đương:

$$P\left(\frac{X-\mu}{\sigma} < u(\alpha)\right) = 1-\alpha$$

$$P\ (X < \mu + \sigma.\ u(\alpha)) = 1-\alpha \qquad (3)$$

Các giá trị $u(\alpha)$ được tìm bằng phương pháp tính gần đúng và dùng máy tính lập trình để tính, do đó chúng đã được tính sẵn và lập thành bảng, chúng ta chỉ việc tra bảng để tìm (bảng I (phụ lục II)).

Hơn nữa, nếu X phân phối chuẩn $(X = N(\mu; \sigma^2))$ thì ta có:

$$P\left(\frac{|X-\mu|}{\sigma} < u\left(\frac{\alpha}{2}\right)\right) = 1-\alpha$$

$$P\left(\mu - \sigma.u\left(\frac{\alpha}{2}\right) < X < \mu + \sigma.u\left(\frac{\alpha}{2}\right)\right) = 1-\alpha \qquad (4)$$

Đặc biệt, nếu $u\left(\frac{\alpha}{2}\right) = 1\ ;\ 2\ ;\ 3$ ta nhận được quy tắc kσ:

(Biết $u\left(\frac{\alpha}{2}\right) = 1$, tra bảng I ta được $\frac{\alpha}{2}$, do đó được α, suy ra $1 - \alpha$ (xem hướng dẫn dưới đây)).

$$P\{|X - \mu| < \sigma\} = P\{\mu - \sigma < X < \mu + \sigma\} = 0{,}682630 \approx 68{,}26\%$$

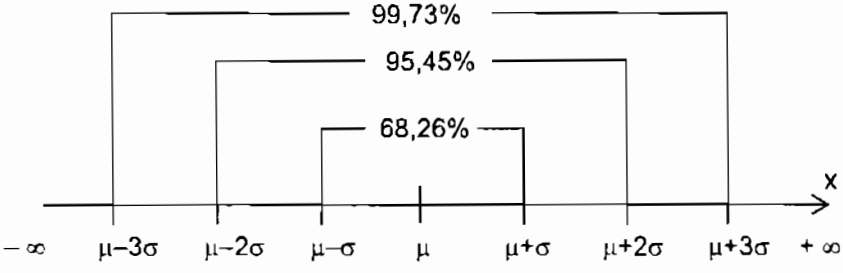
$$P\{|X - \mu| < 2\sigma\} = P\{\mu - 2\sigma < X < \mu + 2\sigma\}$$

$$= 0{,}9545 \approx 95{,}45\% \qquad (5)$$

$$P\{|X - \mu| < 3\sigma\} = P\{\mu - 3\sigma < X < \mu + 3\sigma\}$$

$$= 0{,}9973 \approx 99{,}73\%$$

Nghĩa là với biến ngẫu nhiên chuẩn có giá trị trung bình là μ, độ lệch tiêu chuẩn là σ thì với xác suất 68,26% ta khẳng định biến ngẫu nhiên chuẩn này nhận giá trị trong khoảng $(\mu - \sigma ; \mu + \sigma)$ hay 68,26% giá trị của biến ngẫu nhiên sẽ nằm trong khoảng $(\mu - \sigma ; \mu + \sigma)$, v.v... Ta mô tả hình học như sau:



Minh họa quy tắc kσ hay khoảng giá trị mà biến ngẫu nhiên $N(\mu; \sigma^2)$ nhận và xác suất nhận khoảng đó tương ứng

Rõ ràng ta cũng có: $$P\left(\frac{|X - \mu|}{\sigma} \geq u\left(\frac{\alpha}{2}\right)\right) = \alpha \qquad (6)$$

Thực ra trong các công thức (2), (3), (4), (5), (6) các dấu đẳng thức có dấu bằng hay không có dấu bằng đều như nhau.

Hướng dẫn cách tra bảng I – Phụ lục II:

– Cho giá trị $u(\alpha) \geq 0$ (tức là cho $x \geq 0$). Ta tìm α theo công thức (2) hoặc $(1- \alpha)$ theo công thức (3).

Cột đầu tiên cho ta giá trị của x (phần đơn vị và phần lẻ hàng chục), hàng đầu tiên của bảng cho ta phần lẻ hàng trăm của x.

Từ giá trị u (hoặc x) đã biết ta xác định được hàng và cột tương ứng với u. Giao của cột và hàng cho ta giá trị $1 - \alpha$ (công thức (3)). Từ đó suy ra α. Chẳng hạn $u(\alpha) = 1,25 \rightarrow$ hàng 1,2; cột 0,05 $\rightarrow$ số 0,891350 chính là $1 - \alpha$, do đó $\alpha = 0,108650$. Với u = 1,25 thì tương ứng $1 - \alpha = 0,89135$ và $\alpha = 0,10865$.

– Cho giá trị $u\left(\frac{\alpha}{2}\right)$. Tìm $1 - \alpha$ theo công thức (4).

Ta cũng làm như phần trên. Chẳng hạn biết $u\left(\frac{\alpha}{2}\right) = 1$, tra bảng ta tìm được số 0,841315. Suy ra $\frac{\alpha}{2} = 1 - 0,841315 = 0,158685$. Do đó $\alpha = 2.0,158685 = 0,317370$. Vậy $1 - \alpha = 0,68263$.

Nếu $u\left(\frac{\alpha}{2}\right) = 2 \rightarrow 0,97725 \rightarrow \frac{\alpha}{2} = 1 - 0,97725 = 0,02275$

$\rightarrow \alpha = 2 \,.\, 0,02275 = 0,04550 \rightarrow 1 - \alpha = 0,9545$

Nếu $u\left(\frac{\alpha}{2}\right) = 3 \rightarrow 0,99865 \rightarrow \alpha = 2(1 - 0,99865) = 0,027$

$\rightarrow 1 - \alpha = 0,9973$.

Tức là chúng ta nhận được quy tắc kσ (5).

Như vậy, hai cách tra bảng để tìm α, $1 - \alpha$ theo công thức (2), (3), (4), (5), (6) khi biết $u(\alpha)$ hoặc $u\left(\frac{\alpha}{2}\right)$ là như nhau: Từ giá trị u ta tìm được một con số. Lấy 1 trừ đi con số đó ta được một số tạm gọi là β. Nếu trong công thức (2) hoặc (3) thì giá trị α cần tìm chính là số β đó, còn trong công thức (4), (5), (6) thì giá trị α cần tìm sẽ bằng $2.\beta$.

– Cho α $(0 < \alpha < 0{,}50)$. Tìm $u(\alpha)$ theo công thức (2).

Ta tra bảng theo thứ tự ngược lại với thứ tự trên, tức là cho α ta suy ra $(1 - \alpha)$. Tìm trong bảng I số trùng với $(1 - \alpha)$ hoặc gần với $(1 - \alpha)$ nhất. Từ vị trí này dóng ra hàng ngang và dóng theo cột dọc ta được x. tức là được $u(\alpha)$ cần tìm.

Chẳng hạn, tìm $u(0{,}10) = ?$

Ta có $\alpha = 0{,}10 \rightarrow 1 - \alpha = 0{,}90$

Trong bảng I: số 0,899727 là số gần với 0,90 nhất. Suy ra hàng ngang là 1,2; cột dọc là 0,08. Vậy $u(0{,}10) = 1{,}28$.

Tương tự, ta có $u(0{,}05) = 1{,}65$ (hoặc 1,64 hoặc 1,645);

$u(0{,}025) = 1{,}96$; $u(0{,}01) = 2{,}33$...

(Trong một số tài liệu khác, giá trị $u(\alpha)$ còn được ký hiệu là Z_α).

Nếu cho α để tìm $u\left(\frac{\alpha}{2}\right)$ chúng ta cũng làm như trên, chỉ cần coi $\frac{\alpha}{2}$ là một giá trị α' và đi tìm $u(\alpha')$. (Ở mức độ 30 tiết, chúng ta quan tâm đến cách tra bảng để tìm $u(\alpha)$).

Nếu $u(\alpha) < 0$ hoặc $\alpha > 0{,}50$ chúng ta cũng suy ra được từ bảng I. Nhưng chúng ta dừng lại ở đây và không tìm hiểu sâu hơn nữa.

Như vậy, nếu cho một biến ngẫu nhiên X có phân phối chuẩn với giá trị cụ thể của tham số μ và σ^2, bằng cách sử dụng bảng I ta có thể tính được các tỷ lệ: $(X > a)$; $(X \le b)$; $(a \le X \le b)$, hoặc tìm các mốc a hoặc b khi cho biết các tỷ lệ tương ứng: $P(X > ?) = \alpha$; $P(X < ?) = 1 - \alpha$. Để làm quen với các điều đó và với việc tra bảng, ta xét ví dụ sau. (Với mức độ 30 tiết học bạn đọc có thể bỏ qua).

Ví dụ 3: Gọi X là chỉ số thông minh IQ (Intellgent Quota) của học sinh lứa tuổi 12 đến 15. Giả sử X có phân phối chuẩn: $X = N(85 ; 25)$.

a) Từ giả thiết đã cho ta có được các thông tin gì về chỉ số IQ?

b) Áp dụng quy tắc kơ ta nhận được các thông tin gì.

c) Với xác suất 0,95 có thể nói chỉ số IQ nhỏ hơn giá trị nào?

d) Với xác suất 10% có thể nói chỉ số IQ vượt quá giá trị nào?

e) Tỷ lệ học sinh có chỉ số IQ trên 90 là bao nhiêu %?

f) Tỷ lệ học sinh có chỉ số IQ dưới 87 là bao nhiêu %?

g) Tỷ lệ học sinh có chỉ số IQ thuộc khoảng (78; 92) là bao nhiêu %?

Giải:

a) Theo giả thiết: $X = N(85 ; 25)$, nghĩa là chỉ số thông minh X có phân phối chuẩn với $EX = 85$; $DX = 25$. Vậy chỉ số thông minh trung bình của học sinh lứa tuổi 12 – 15 là 85; độ lệch tiêu chuẩn là 5.

b) Theo quy tắc kơ ta có:

$P\{85 - 5 < X < 85 + 5\} = P\{80 < X < 90\} = 68,26\%$

$P\{85 - 2.5 < X < 85 + 2.5\} = P\{75 < X < 95\} = 95,45\%$

$P\{85 - 3.5 < X < 85 + 3.5\} = P\{70 < X < 100\} = 99,73\%$

Tức là tỷ lệ học sinh có chỉ số IQ trong khoảng (80 ; 90) là 68,26% ; tỷ lệ học sinh có chỉ số IQ trong khoảng (75 ; 95) là 95,45%; còn tỷ lệ học sinh có chỉ số thông minh trong khoảng (70 ; 100) là 99,73%.

c) Theo công thức (3): $P(X < \mu + \sigma.u(\alpha)) = 1 - \alpha$, cho nên với xác suất 95%, tức $1 - \alpha = 0,95 \rightarrow \alpha = 0,05 \rightarrow u(0,05)$. Tra bảng ta được $u(0,05) = 1,65$.

Vậy $P(X < 85 + 5.1,65) = 0,95$.

$\rightarrow P(X < 93,25) = 0,95$. Tức là, với xác suất 0,95 chỉ số IQ của học sinh nhỏ hơn 93,25.

d) Tương tự, từ công thức (2): $P(X \geq \mu + \sigma.u(\alpha)) = \alpha$. Áp dụng với $\alpha = 0,10$ suy ra giá trị $u(\alpha)$ là $u(0,10)$. Tra bảng ta được $u(0,10) = 1,28$. Do đó:

$P(X \geq 85 + 5.1,28) = 0,10$

$P(X \geq 91,40) = 0,10$

Với xác suất 10% ta có thể kết luận chỉ số thông minh của học sinh trên mức 91,40.

e) Tỷ lệ học sinh có chỉ số IQ trên 90 chính là $P(X \geq 90)$. Theo (2) ta có:

$P(X \geq 90) = P(X \geq 85 + 5) = P(X \geq 85 + 5.1) = \alpha$.

Theo công thức (2) ta thấy $u(\alpha) = 1$. Tra bảng ta được $1 - \alpha = 0,841315$.

Suy ra $\alpha = 0,158685$. Vậy $P(X \geq 90) = 0,158685 \approx 16\%$.

f) Ta lại có:

$$P(X < 87) = P(X < 85 + 2) = P(X < 85 + 5.\frac{2}{5}) = 1 - \alpha.$$

Theo công thức (3) ta thấy $u(\alpha) = +\frac{2}{5} = +0,4$. Tra bảng ta

được $1 - \alpha = 0,655422$. Vậy $P(X < 87) = 0,65522 \approx 65,54\%$.

Tỷ lệ học sinh có chỉ số IQ dưới 87 là 65,54%.

g) Để tính tỷ lệ học sinh có IQ thuộc khoảng (78 ; 92) chúng ta dùng công thức (4). Ta có:

$$P(78 < X < 92) = P(85 - 7 < X < 85 + 7)$$

$$= P(85 - 5.\frac{7}{5} < X < 85 + 5.\frac{7}{5})$$

Suy ra $u\left(\frac{\alpha}{2}\right) = \frac{7}{5} = 1,4$. Ứng với $u = 1,4$ ta tìm được số $0,919213. \rightarrow \alpha = 2.(1 - 0,919213) = 0,161574$.

Vậy $1 - \alpha = 0,838426 \approx 83,84\%$

Tỷ lệ học sinh có IQ $\in$ (78 ; 92) là 83,84%.

9. Phân phối khi bình phương, ký hiệu χ^2_k

Liên quan đến phân phối này chúng ta chỉ dùng các giá trị $\chi^2_k(\alpha)$ được xác định từ: P{biến ngẫu nhiên khi bình phương với tham số k nhận giá trị lớn hơn $\chi^2_k(\alpha)$} = α. Các giá trị $\chi^2_k(\alpha)$ được tính sẵn và lập thành bảng II (Phụ lục II). Để tìm $\chi^2_k(\alpha)$ chúng ta đối chiếu hàng ngang theo k, cột dọc theo α. Giao điểm chính là giá trị $\chi^2_k(\alpha)$ cần tìm.

10. Phân phối Student

Liên quan đến phân phối này chúng ta sẽ dùng tới các giá trị $t_k(\alpha)$ được xác định từ P{biến ngẫu nhiên Student với tham số k nhận giá trị lớn hơn $t_k(\alpha)$} = α. Các giá trị $t_k(\alpha)$ được tra từ bảng III (Phụ lục II). Cách tra bảng cũng giống như tra bảng χ^2.

Ký hiệu t_k là biến ngẫu nhiên có phân phối Student với tham số k. Khi đó ta có:

$$P\left\{|t_k| < t_k\left(\frac{\alpha}{2}\right)\right\} = P\left\{-t_k\left(\frac{\alpha}{2}\right) < |t_k| < t_k\left(\frac{\alpha}{2}\right)\right\} = 1 - \alpha \quad (7)$$

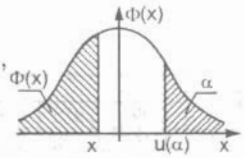
$$P\left\{|t_k| \geq t_k\left(\frac{\alpha}{2}\right)\right\} = \alpha \quad (8)$$

Trong đó $t_k\left(\frac{\alpha}{2}\right)$ được xác định tương tự như $t_k(\alpha)$, nghĩa là $P\left\{t_k > t_k\left(\frac{\alpha}{2}\right)\right\} = \frac{\alpha}{2}$ và được tìm từ bảng III.

PHỤ LỤC II

CÁC BẢNG SỐ

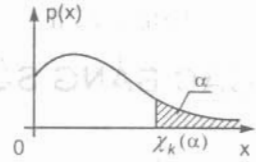
Bảng 1. Bảng giá trị $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du = \Phi(x)$,



$0 \le x \le 3$

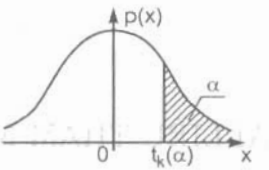
x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	,500000	,503909	,507978	,514966	,545953	,519938	,523922	,527903	,531881	,535856
0,1	,539828	,513795	,517758	,551717	,555670	,559618	,563560	,567495	,571424	,575345
0,2	,579260	,583166	,587061	,590954	,597835	698706	,602658	,606120	,610261	,614092
0,3	,617911	,621720	,625616	,629300	,633072	636834	,640576	,644309	,648027	,651732
0,4	,655422	,659097	,662757	,66102	,670031	,673645	,677242	,680622	,684386	,687933
0,5	,694462	,691971	,698186	,702914	,705102	,708840	,712260	,715661	,719013	,722105
0,6	,725717	,729669	,732371	,736653	,738914	,712154	,745373	,748571	,751718	,751903
0,7	,758036	,767148	,764238	,797305	,740350	,773373	,776373	,779350	,782305	,735236
0,8	,788145	,791030	,793892	,796731	,799546	,802338	,805160	,807850	,810570	,893267
0,9	,815910	,818589	,721214	,823814	,826391	,828944	,831472	,833977	,836457	,838913
1,0	,841315	,843752	,746136	,848495	,850830	,853141	,855428	,857690	,859929	,862143
1,1	,861331	,866570	,868843	,870763	,872857	,874928	,876976	,879000	,881000	,882977
1,2	,881930	,886861	,888768	,890654	,897512	,891350	,896165	,897958	,899727	,901457
1,3	,903200	,901902	,906582	,908241	,909877	,914492	,913085	,914656	,916207	,917736
1,4	,919213	,920730	,922196	,923642	,925066	,926477	,927855	,929210	,930563	,931889
1,5	,933193	,934478	,935744	,936922	,938220	,939429	,940620	,941792	,942917	,944083
1,6	,945201	,956301	,947384	,948149	,949497	,950528	,951543	,952210	,953521	,954486
1,7	,955131	,956367	,957284	,958185	,959070	,959941	,960796	,961636	,962462	,963373
1,8	,961070	,961852	,965620	,966375	,967116	,967843	,968557	,969258	,969946	,970621
1,9	,971283	,971933	,972571	,973107	,973810	,974412	,975002	,975581	,976138	,976701
2,0	,977250	,977781	,978308	,978822	,979325	,979818	,980301	,980774	,981237	,981691
2,1	982136	,982571	,982997	,983414	,983823	,984222	,984614	,984997	,985377	,985738
2,2	986097	,986447	,986791	,987126	,987454	,987776	,988089	,988396	,988696	,988989
2,3	989276	,989556	,989830	,990097	,990358	,990613	,990762	,991106	,991314	,991576
2,4	,99180	,992024	,992240	,992451	,992656	,992857	,993053	,993244	,994431	,993613
2,5	,993790	,993963	,994132	,994297	,994457	,994611	,991766	,994915	,995060	,995201
2,6	,995339	,995173	,995604	,995731	,995855	,995975	,996093	,996007	,996319	,996127
2,7	,996533	,996636	,996736	,996833	,996928	,997920	,997110	,997197	,997282	,997365
2,8	,997415	,997523	,997599	,997673	,997714	,997814	,997882	,997948	,998012	,998074
2,9	,998134	,998183	,998250	,998305	,998359	,998411	,998462	,998511	,998559	,998605
	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
3,0	,998650	,999032	,999313	,999517	,999663	,999767	,999841	,999892	,999928	,999952

Bảng 2. Tìm $\chi^2_k(\alpha)$ theo bậc tự do k và α



k	α											
	0,99	0,975	0,95	0,90	0,70	0,50	0,30	0,10	0,05	0,025	0,01	0,001
1	,02157	,03982	,0393	,0158	,018	,455	1,07	2,71	3,84	5,02	6,62	10,8
2	,0201	,0506	,103	,214	,713	1,39	2,41	4,64	5,99	7,38	9,24	13,8
3	,0145	,216	,352	,581	1,12	2,37	3,67	6,25	7,81	9,15	11,3	16,3
4	,297	,184	,714	1,06	2,19	3,36	4,88	7,78	9,49	11,1	13,3	18,5
5	,554	,834	1,15	1,64	3,00	4,35	6,06	9,24	11,1	12,3	15,1	20,5
6	,872	1,21	1,64	2,20	3,83	5,35	7,23	10,6	12,6	14,1	16,8	22,5
7	1,24	1,69	2,71	2,83	4,67	6,35	8,38	12,0	14,1	16,0	18,5	24,3
8	1,65	2,18	2,73	3,49	5,53	7,34	9,52	13,4	15,5	17,5	20,1	26,1
9	2,09	2,70	3,33	4,17	6,39	8,34	10,7	14,7	16,9	19,0	21,7	27,9
10	2,56	3,25	3,94	4,57	7,27	9,34	11,8	16,0	18,3	20,5	23,2	29,6
11	3,05	3,82	4,57	5,58	8,15	10,3	12,9	17,3	19,7	21,9	24,7	31,3
12	3,57	4,10	5,23	6,30	9,03	11,3	14,0	18,5	21,0	23,3	26,2	32,9
13	4,14	5,01	5,89	7,04	9,93	12,3	15,1	19,8	22,4	24,7	27,7	34,5
14	4,66	5,63	6,57	7,79	10,8	13,3	16,2	21,1	23,7	26,1	29,1	36,1
15	5,23	6,26	7,26	8,55	11,7	14,3	17,3	22,3	25,0	27,5	30,6	37,7
16	5,81	6,91	7,96	9,31	12,6	15,3	18,4	23,5	26,3	28,8	32,0	39,3
17	6,14	7,56	8,67	10,1	13,5	16,3	19,5	34,4	27,6	30,2	33,4	40,8
18	7,01	8,23	9,30	10,9	14,4	17,3	20,6	26,0	28,9	31,5	34,8	42,3
19	7,63	8,94	10,1	11,7	15,4	18,3	21,7	27,2	30,1	32,9	36,2	43,8
20	8,26	9,59	10,9	12,4	16,3	19,3	22,8	28,1	31,4	34,2	37,6	45,3
21	8,90	10,3	11,6	13,2	17,2	20,3	23,9	29,6	32,7	35,5	38,9	46,8
22	9,54	11,0	12,3	14,0	18,1	21,3	24,9	30,8	33,9	36,8	40,3	48,3
23	10,2	11,7	13,1	14,8	19,0	22,3	26,0	32,0	35,2	38,1	41,6	49,7
24	10,9	12,4	13,8	15,7	19,9	23,3	27,1	33,2	36,4	39,4	43,0	51,2
25	11,5	13,1	14,6	16,5	20,9	24,3	28,2	34,4	37,7	40,6	44,3	52,6
26	12,2	13,8	15,4	17,3	21,8	25,3	29,2	35,6	38,9	41,9	45,6	54,1
27	12,9	14,6	16,2	18,1	22,7	26,3	30,3	36,7	40,1	43,2	47,0	55,5
28	13,6	15,3	16,9	18,9	23,6	27,3	31,4	37,9	41,3	44,5	48,3	56,9
29	14,3	16,0	17,7	19,8	24,6	28,3	32,5	39,1	42,6	45,7	49,6	58,3
30	15,0	16,8	18,5	20,6	25,5	29,3	33,5	40,3	43,8	47,0	50,9	59,7
40	22,2	21,4	26,5	29,1	34,9	39,3	44,2	51,8	55,8	59,3	63,7	73,1
50	29,7	32,4	34,8	37,7	44,3	49,3	54,7	63,2	67,5	71,4	76,2	86,7
60	35,5	40,5	43,2	46,5	53,8	59,3	65,2	74,4	79,1	83,3	88,4	99,6
70	45,1	48,8	51,7	55,3	63,3	69,3	75,1	85,5	90,5	95,0	100,4	112,3
80	53,5	57,2	60,4	61,3	72,9	79,3	86,1	96,6	101,9	106,6	112,3	124,8
90	61,8	65,6	69,1	73,3	82,5	89,3	96,5	107,6	113,1	118,1	124,1	137,2
100	70,1	74,2	77,9	82,4	92,1	99,3	106,9	118,5	124,3	129,6	135,8	149,1

Bảng 3. Giá trị $t_k(\alpha)$ của phân bố t (phân bố Student)



Bậc tự do k	Mức ý nghĩa α						Bậc tự do k	Mức ý nghĩa α					
	0,05	0,025	0,01	0,005	0,001	0,0005		0,05	0,025	0,01	0,005	0,001	0,0005
1	6,34	12,71	31,82	63,66	318,34	636,62	18	1,73	2,10	2,55	2,88	3,61	3,92
2	2,92	4,30	6,97	9,92	22,33	31,06	19	1,73	2,09	2,54	2,86	3,58	3,88
3	2,35	3,18	4,51	5,84	10,21	12,92	20	1,73	2,09	2,53	2,85	3,55	3,85
4	2,13	2,78	3,75	4,60	7,17	8,61	21	1,72	2,08	2,52	2,83	3,53	3,82
5	2,01	2,57	3,37	4,03	5,89	6,87	22	1,72	2,07	2,51	2,82	3,51	3,79
6	1,91	2,45	3,14	3,71	5,21	5,96	23	1,71	2,07	2,50	2,81	3,49	3,77
7	1,89	2,36	3,00	3,50	4,79	5,41	24	1,71	2,06	2,49	2,80	3,47	3,74
8	1,86	2,31	2,90	3,36	4,50	5,04	25	1,71	2,06	2,49	2,79	3,45	3,72
9	1,83	2,26	2,82	3,25	4,30	4,78	26	1,71	2,06	2,48	2,78	3,44	3,71
10	1,81	2,23	2,76	3,17	4,11	4,59	27	1,70	2,05	2,47	2,77	3,42	3,69
11	1,80	2,20	2,72	3,11	4,03	4,44	28	1,70	2,05	2,47	2,76	3,41	3,66
12	1,78	2,18	2,68	3,05	3,93	4,32	29	1,70	2,05	2,46	2,76	3,40	3,66
13	1,77	2,16	2,65	3,01	3,85	4,22	30	1,70	2,01	2,46	2,75	3,39	3,65
14	1,76	2,14	2,62	2,98	3,79	4,11	40	1,68	2,02	2,42	2,70	3,31	3,55
15	1,75	2,13	2,60	2,95	3,73	4,07	60	1,67	2,00	2,39	2,66	3,23	3,46
16	1,75	2,12	2,58	2,92	3,69	4,01	120	1,66	1,98	2,36	2,62	3,17	3,37
17	1,71	2,11	2,57	2,90	3,65	3,96	∞	1,61	1,96	2,33	2,58	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005		0,05	0,025	0,01	0,005	0,001	0,0005
	Mức ý nghĩa α							Mức ý nghĩa α					

TÀI LIỆU THAM KHẢO

1. Đào Hữu Hồ, *Xác suất thống kê*,
NXB Đại học Quốc gia Hà Nội, In lần thứ 9, năm 2006.
2. Lincoln L.Chao, *Statistics: Methods and Analyses*
International Student Edition, 1969.

MỤC LỤC

	Trang
Lời nói đầu	3
Chương I: **Một số khái niệm và kết quả cơ bản của xác suất**	5
I.1. Giải tích tổ hợp	5
I.2. Phép thử và biến cố	7
I.3. Định nghĩa xác suất	8
I.4. Công thức xác suất của tổng và tích hai biến cố	16
I.5. Dãy phép thử Bernoulli	20
I.5.1. Định nghĩa	20
I.5.2. Xác suất $P_n(m; p)$	21
I.5.3. Số có khả năng nhất	22
I.6. Biến ngẫu nhiên	24
I.6.1. Định nghĩa	24
I.6.2. Biến ngẫu nhiên rời rạc	25
I.6.3. Biến ngẫu nhiên liên tục	29
I.7. Hàm phân phối	30
I.7.1. Định nghĩa	30
I.7.2. Tính chất	30
I.8. Các số đặc trưng của biến ngẫu nhiên	34
I.8.1. Kỳ vọng (giá trị trung bình)	34
I.8.2. Phương sai	36
I.8.3. Mode	37
I.8.4. Median và phân vị	37

I.9. Một vài phân phối cần dùng 38
I.9.1. Phân phối nhị thức B(n; p) 39
I.9.2. Phân phối siêu bội (siêu hình học) 41
I.9.3. Phân phối chuẩn $N(\mu; \sigma^2)$ 42
I.9.4 Phân phối khi bình phương χ_k^2 46
I.9.5. Phân phối Student t_k 47
Bài tập chương I 49
Chương II: **Thống kê xã hội** 54
II.1. Giới thiệu bài toán 54
II.2. Lý thuyết mẫu 56
II.2.1. Một vài phương pháp lấy mẫu đơn giản 56
II.2.2. Mẫu ngẫu nhiên 60
II.2.3. Cách thu gọn và biểu diễn số liệu 63
II.2.4. Các đặc trưng mẫu 75
II.2.5. Cách tính $\bar{X}$ và s^2 77
II.2.6. Sai số trong lấy mẫu 82
II.3. Một vài ước lượng đơn giản 85
II.3.1. Ước lượng điểm cho kỳ vọng, Median, Mode, phương sai và xác suất 85
II.3.2. Ước lượng khoảng cho kỳ vọng và xác suất (cho giá trị trung bình và tỷ lệ) 92
II.4. Một số bài toán kiểm định giả thiết đơn giản 101
II.4.1. Đặt bài toán 101
II.4.2. Kiểm định giả thiết về giá trị trung bình 104
II.4.3. Kiểm định giả thiết về tỷ lệ 109
II.4.4. So sánh hai giá trị trung bình 111
II.4.5. So sánh hai tỷ lệ (hai xác suất) 124
II.4.6. Tiêu chuẩn phù hợp χ^2 126
II.4.7. Kiểm tra tính độc lập 129

II.4.8. So sánh nhiều tỷ lệ 133
II.5. Tương quan và hồi quy đơn 135
II.5.1. Hệ số tương quan 136
II.5.2. Đường hồi quy bình phương trung bình tuyến tính 138
Bài tập chương II 144
Hướng dẫn giải và đáp số 159
Phụ lục I: Một số khái niệm của xác suất 180
Phụ lục II: Các bảng số 201
Tài liệu tham khảo 204

Chịu trách nhiệm xuất bản:

Chủ tịch HĐQT kiêm Tổng Giám đốc NGÔ TRẦN ÁI
Phó Tổng Giám đốc kiêm Tổng biên tập NGUYỄN QUÝ THAO

Tổ chức bản thảo và chịu trách nhiệm nội dung:
Chủ tịch HĐQT kiêm Giám đốc CTCP Sách ĐH–DN

TRẦN NHẬT TÂN

Biên tập và sửa bản in:

HOÀNG THỊ QUY

Trình bày bìa:
HOÀNG MẠNH DỨA

Chế bản:
ĐINH XUÂN DŨNG

GIÁO TRÌNH THỐNG KÊ XÃ HỘI HỌC

Mã số: 7L189M7 - DAI

In 2.000 bản (QĐ 30), khổ 14.5 x 20.5 cm, tại Công ty In - Thương mại TTXVN
70/342 Khương Đình - Hạ Đình - Thanh Xuân - Hà Nội
Số xuất bản: 17-2007/CXB/101-2217/GD
In xong và nộp lưu chiểu tháng 5 năm 2007

CÔNG TY CỔ PHẦN SÁCH ĐẠI HỌC - DẠY NGHỀ
HEVOBCO
25 HÀN THUYÊN – HÀ NỘI
Website : www.hevobco.com.vn

CÙNG MỘT TÁC GIẢ

1. **Thống kê Toán học** (Đồng tác giả). NXB Đại học và Trung học chuyên nghiệp, 1984. NXB Đại học Quốc gia Hà Nội, 2004.
2. **Xác suất Thống kê**, NXB Đại học Quốc gia Hà Nội.
   In lần đầu. 1996. In lần thứ 9, 2006.
3. **Thống kê xã hội học**, NXB Đại học Quốc gia Hà Nội.
   In lần đầu. 1996. In lần thứ 7, 2006.
4. **Thống kê sinh học** (Đồng tác giả), NXB Khoa học và Kỹ thuật, Hà Nội 1999, 2001.
5. **Xử lý số liệu bằng Thống kê Toán học trên máy tính** (Đồng tác giả), NXB Đại học Quốc gia Hà Nội – In lần đầu, 2000. In lần thứ 3, 2004.
6. **Xác suất Thống kê - Chương trình Giáo dục Đại học 1997, 1998**, NXB Khoa học và Kỹ thuật. Hà Nội, 2002.
7. **Hướng dẫn giải các bài toán xác suất thống kê**, NXB Đại học Quốc gia Hà Nội. 2004, 2006. Sách được giải "Sách hay" – Giải thưởng sách Việt Nam lần thứ I (tháng 5/2006).

Ban đọc có thể mua tại các Công ti Sách - Thiết bị trường học ở các địa phương hoặc các Cửa hàng của Nhà xuất bản Giáo dục :

Tại Hà Nội : 25 Hàn Thuyên, 187B Giảng Võ, 232 Tây Sơn, 23 Tràng Tiền.

Tại Đà Nẵng : So 15 Nguyễn Chí Thanh. Số 62 Nguyễn Chí Thanh.

Tại Thành phố Hồ Chí Minh : 104 Mai Thị Lưu – Quận 1, Cửa hàng 451B - 453. Hai Bà Trưng – Quận 3; 240 Trần Bình Trọng – Quận 5.

Tại Thành phố Cần Thơ : Số 5/5. đường 30/4.

Website : www.nxbgd.com.vn

8 934980 763483

Giá: 17.500 đ